

Business and pleasure can mix

The British government should not miss an opportunity to show its support for an innovative scheme from the Wellcome Trust to help take genomics research from the lab to the market.

In the next few months, Britain's deputy prime minister and environment secretary, John Prescott, will face a particularly tricky decision. He has to decide, on the basis of evidence presented last week at a public inquiry, whether to allow the Wellcome Trust, the world's largest medical research charity, to extend its genome-research campus at Hinxton, a rural location ten miles south of Cambridge, to provide 'incubator' space for spin-off companies (see pages 355–356).

The trust is not planning an ordinary research park. It wants to create an environment where the boundaries between fundamental research and its commercial applications — and between work and leisure — are blurred. It has a vision of ground-breaking ideas in genomics-based health care emerging from both planned and chance interactions between scientists and entrepreneurs.

The proposed extension includes shared conference rooms and coffee bars, a cricket field, and facilities for bird-watching, fishing and other leisure pursuits. Equally important, potential entrepreneurs will have access to purpose-built accommodation on terms the market will find hard to match. But plans for the extension have been rejected by the local council on the grounds that any new development on agricultural land away from public transport contravenes government policies to protect the countryside and discourage journeys by car.

The trust has tried to meet the content, if not the spirit, of such

objections. It has promised to put in place measures to meet most of the local authority's targets to reduce the number of car journeys. It has also promised to pay for extra bus services and to penalize cars parked on the extension without authorization. And it insists that there will be no commercial manufacturing on the site, and that companies that survive more than six years will be asked to move elsewhere.

Prescott, therefore, has fewer reasons than initially to refuse the trust's application. He may be tempted to do so from Wellcome's less-than-convincing defence of its claim that companies housed within walking distance of research laboratories perform better than those further away. This will be a tricky issue for the deputy prime minister, as there is little empirical evidence to support the case for 'co-location'. Supportive words from those who have already enjoyed such arrangements are not the same as hard data on its relative effectiveness.

But an absence of hard numbers should not be a reason to reject what is an innovative — and, in its revised form, a much more 'environmentally friendly' — scheme, with the potential to contribute significantly to our treatment of disease. The trust has shown it is willing to accommodate local concerns, even if it needs to listen more and work harder to allay negative perceptions of itself. It is time for South Cambridgeshire District Council and local residents to join in the spirit of compromise and lift their objections. □

Pulling together in Latin America

Greater pan-regional scientific collaboration could raise Latin American science to the next level of excellence.

Latin America, whatever the recent economic travails of Brazil, is well placed to strengthen its already growing contribution to the world of science (as a special supplement to this issue demonstrates). Admittedly, day-to-day circumstances remain difficult for scientists in the region. Progress sometimes seems to be faltering, with governments continually introducing new schemes to support science — and then failing to find the money to follow them through.

But the international influence of Latin American scientists is steadily expanding, and there is every indication that it will continue to do so. Certainly, the overall trajectory, measured, for want of a better yardstick, by the proportion of papers from the region published in international journals, points firmly upwards.

As scientists from the region's widely disparate countries jostle to both compete and collaborate with their better-supported peers in the United States and Europe, however, it is becoming increasingly important for them to recognize how much they can learn from the experience of their neighbours. Most would, indeed, like to work more closely with their peers in neighbouring countries. At present, though, there is little encouragement to do so.

Most of these researchers trained in the United States or in Europe, and are inevitably keen to maintain strong connections with the wealthy research institutions of the so-called developed world. It is certainly far easier to work within an international collaboration if at least one end of that collaboration is adequately funded.

There are plenty of isolated examples of scientific collaboration

between the countries of the region. At present, however, there is no overall strategy to pool resources in a way that would strengthen the region's scientific base. Many officials — and some scientists — continue to believe that the only worthwhile activity in their particular area of interest lies far to the north.

But research collaboration between north and south is inevitably asymmetrical, with limits on its usefulness to the weaker party. As Latin American science grows in scope and confidence, the region's relatively small scientific communities will greatly increase their ability to compete internationally by combining forces to create the critical mass needed to produce high-quality science in their own region.

In Europe, scientific collaboration has played a significant role in boosting national scientific capabilities, particularly when faced with competition from the United States. The European Laboratory for Particle Physics (CERN) and the European Molecular Biology Laboratory are just two examples of what has been achieved. Latin America now faces a similar challenge.

A number of proposals under consideration, such as the Ibero-American Molecular Biology Organization's plans for a centre of excellence and the World Bank's Millennium Institutes, could catalyse a regional response. President Eduardo Frei of Chile has tried — apparently with little success — to place research collaboration on the agenda at summits of the region's leaders. If they want to build competitive economies, the region's main powers — Brazil, Argentina and Mexico — must address the need for pan-regional collaboration. □



Wellcome wages battle to house companies next to genome labs

[CAMBRIDGE, ENGLAND] Britain's Wellcome Trust is locked in a bitter battle with local planning authorities over a proposed £100 million (US\$162 million) commercial extension to its Genome Campus in Hinxton, Cambridgeshire.

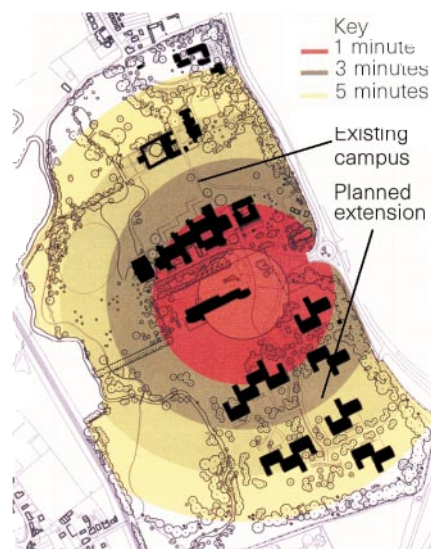
At a public inquiry last week, the local authority confirmed its opposition to the extension, which the trust wants to provide space for small biotechnology companies involved in exploiting the campus's work on genome sequencing.

The trust argues that it is essential for these companies to be next to the campus to maximize their chance of success. But the local authority says the development would contravene government guidelines on the use of agricultural land and transport policy.

The final decision on the plan now rests with the deputy prime minister and environment secretary, John Prescott. His judgement will reflect the extent to which the government's pledge to support knowledge-based businesses is compatible with its desires to protect the countryside and limit the use of cars.

The trust wants to provide space for up to 600 people in genome-based businesses within walking distance of its Hinxton campus, 10 miles from the city of Cambridge and next to the villages of Hinxton and Ickleton.

It claims that, without the 40,000 square metres of extra space, Britain will lose a valuable opportunity to commercialize



Short walk: the new extension would be a few minutes from Hinxton's Genome Campus (top).

research spinning out of the genome campus's three laboratories: the Sanger Centre for gene sequencing, the European Bioinformatics Institute, and the Human Genome Mapping Project Resource Centre.

But South Cambridgeshire District Council says that space for 'spin-out' companies already exists on the district's many science parks. And it argues that further construction on the Hinxton site contravenes government policies to restrict new building

on agricultural land and in places that are not adequately served by public transport.

The inquiry was chaired by a government planning inspector, who will advise Prescott of his decision in the coming weeks. The inspector has already advised against the extension once. But, given the potential significance of the scheme to the UK economy, Prescott gave Wellcome a second chance.

The trust was asked to justify why it needs the extra space, and to explain how it intends to match the council's green transport target of a 50 per cent reduction in car journeys when the extension opens in 2001.

Prescott has in principle accepted the need for a smaller 'innovation centre' — a building housing companies up to three years old. But the trust wants additional space to house 'nursery'-stage companies up to six years old that have outgrown the centre. It estimates that demand for premises will come from 28 companies a year during the first two years.

During the public inquiry's three days, the trust relied heavily on endorsements for its plan from leading bioscientists with experience of research commercialization. It also unveiled its 'green travel plan', which includes measures to encourage employee car sharing, extra buses between the campus and nearby cities and towns, and on-site parking restrictions.

The council reluctantly acknowledged that these measures could contribute substantially towards meeting its green travel targets. But it refused to concede the Wellcome Trust's main contention that businesses that rely on results from the Human Genome Project need to be within walking distance of the Hinxton campus.

Michael Morgan, chief executive of the genome campus, says there are two main ideas behind the plan to locate biotech companies alongside the research workers. One is to provide scientists and entrepreneurs with formal and informal meeting places, such as shared conference and coffee rooms, sports and social facilities, to help generate ideas. The second is to make it easier for scientists who set up companies to maintain links with the laboratories in which they had worked.

But Simon Bird, the barrister for South Cambridgeshire District Council, pointed out that the trust was unable to produce substantial quantitative evidence that spin-out companies within walking distance of research laboratories are more likely to succeed than those located within driving distance. This point was also made by witnesses

Local residents unimpressed by growth plan

[CAMBRIDGE, ENGLAND] The Wellcome Trust's proposed Genome Campus extension is being opposed by the residents of the nearby villages of Hinxton and Ickleton — combined population of fewer than 1,000 — who see the plan as the thin end of a wedge.

Part of their opposition stems from a fear that the trust is planning an entire 'genome village', which they believe will destroy their rural way of life. Some also attack the glossy corporate image that the trust has developed, arguing that, although it may be successful in influencing government, it is out of touch with local feeling.

Michael Morgan, chief executive of the Hinxton campus, acknowledges this perception, and agrees that the trust could improve its links with local people. He is keen to dispel any perceptions that the Wellcome Trust, which prides itself on being the world's largest medical research charity, is a commercial organization.

But councillors and residents — who include retired academics, teachers and business people — say they do not want to be seen as anti-science. Many of those who spoke at last week's inquiry praised the government's efforts to

support knowledge-driven economic growth.

Some refuted the idea of placing business units within walking distance of laboratories, for example, by pointing out that there are few examples of this in the United States. One retired scientist who later went into management says the idea that academics and entrepreneurs will happily share information during coffee breaks will not work in the long run.

"It might work when there are few companies on campus," he says. "But when there are 50, there will be too much competition for that sort of thing."

E.M.



Existing buildings on the Hinxtan campus, home of Britain's role in the Human Genome Project.

who supported the council's view. One such witness was Roger Quince, managing director of a nearby science park, Granta Park, and a founder of the technology management consultancy Segal, Quince, Wicksteed. He said the plan presupposes that research at the campus will lead directly to commercialization — the so-called 'linear model'.

Quince pointed out that individuals who were located elsewhere, such as at research departments at the University of Cambridge, hospitals and other businesses, were equally likely to contribute to the development of drugs from genomics.

Despite the absence of quantitative data, the trust had obtained letters of support from senior scientists representing science parks in Israel and the United States, as well as from major pharmaceutical companies.

Support also came from PPL Therapeutics, which in 1992 moved from premises at the University of Edinburgh to offices within 200 metres of the Roslin Institute, home of Dolly, the first cloned sheep. "I am convinced that our extremely successful participation, in what turned out to be a major scientific breakthrough, would not have occurred had we still been five miles down the road in Edinburgh University," wrote Alan Coleman, PPL's research director.

Ehsan Masood

New Howard Hughes head seeks bioinformatics boost

[WASHINGTON] The Howard Hughes Medical Institute last week named biochemist Thomas Cech as its next president.

Cech won the Nobel Prize in Chemistry in 1989 for his work on the biocatalytic function of ribonucleic acid. He intends to maintain his laboratory at the University of Colorado at Boulder while running the United States' largest research philanthropy from its headquarters at Chevy Chase, Maryland.

"My two great passions have been biomedical research and science education," Cech says. "This is an opportunity to get involved in these issues at the highest level."

The 51-year-old scientist hopes to see Hughes taking a more active role in applying information technology to biology. "Computing is revolutionizing biology, and we could be at the forefront of bioinformatics," he says. "The institute could support investments in bioinformatics and, for example, set up regional centres" for computing.

Hughes currently spends most of its \$420 million science programme on directly supporting 318 US investigators, among them many of the country's most accomplished life scientists, including Cech. It also spends \$100 million each year on a grants programme that supports the reform of life-science teaching at US universities, as well as researchers outside the United States.

Cech, who has always taught chemistry to first-year students at his own university, is very enthusiastic about the university teaching programme. "It's a real American success story," he says. "Everywhere I go, institutes have been using this money to redefine the way they educate science students."

He seems less sure about the international programmes that support scientists in



Cech: a passion for science education.

Eastern Europe, Latin America and Canada. "I need some time to talk to people" about the programmes, he says, "and to find out what administrative burden they place on Hughes." Cech also questions why the programmes only support scientists in some parts of the world.

While he admires the UK-based Wellcome Trust, the only medical research philanthropy of comparable size, Cech doesn't expect to follow its example by influencing government research policy.

"Lobbying the government is not on my list of activities," he says. Hughes may, however, raise its profile by becoming more involved in public discussion of the ethical, legal and other issues surrounding science.

The Howard Hughes Medical Institute was set up by the billionaire aviator as a tax dodge in 1953, but was reconstituted after his death. Independent trustees gained control in 1984, sold its main asset, Hughes Aircraft, to General Motors, and reached an agreement with tax authorities that requires it to spend at least 3.5 per cent of its vast endowment on medical research.

The stock-market boom has taken the value of that endowment to more than \$11 billion. Cech will probably become the world's best-paid science administrator next January. His predecessor Purnell Choppin, president since 1987, earned more than \$600,000 a year.

Colin Macilwain

Scientists welcome Prodi as European Commission president

[MUNICH] Last week's appointment of former Italian prime minister Romano Prodi as president of the European Commission, following the resignation of his predecessor Jacques Santer, could be good news for Europe's scientific community.

Prodi, a professor of economics at the University of Bologna, is a firm believer in the importance of science to a strong, stable economy (see *Nature* 375, 620; 1995). As prime minister he oversaw the start of a major review of Italy's science institutions.

Glauco Tocchini-Valentini, director of the Italian National Research Council's Institute of Cell Biology at Monterotondo, says Prodi has always been very sensitive to science.

Enric Banda, general secretary of the

European Science Foundation, calls Prodi "a federalist who believes in promoting cooperation between institutions". He believes that Prodi will be sympathetic to the concept of significant levels of basic research being carried out at European level.

"Fundamental research is certainly seen as a federal issue in the United States," Banda points out. In contrast, basic research has had to fight to maintain its small presence in the European Union's fifth Framework programme (FP5), as it is considered primarily the responsibility of individual member states.

"I'm sure that his appointment as head of the commission must be good for science in the medium term — even if science

becomes temporarily swallowed up by other political priorities," says Banda. Such was the fate of the Italian science review, which is only now approaching completion.

Prodi's choice of research commissioner, which should become clear by early summer, will also be important. But as he puts together his new commission over the next month, European scientists need not fear that FP5 project funding will be delayed.

The research commission, foreseeing the possible departure of the commissioners — who resigned *en bloc* following a critical report — ensured that any critical decisions on individual FP5 programmes were taken by early March. Calls for proposals for each of them have now been issued. **Alison Abbott**

Wyoming physics faculty faces closure

[WASHINGTON] The University of Wyoming is considering closing its physics and astronomy department, a move that would make it the only state-run university in the United States not to award degrees in physics.

The plan is still in draft form and is not scheduled to go to the university's board of trustees for approval until July. But physics department chairman Paul Johnson acknowledges that his programme, which has seen a heavy fall in student numbers over the past few years, is facing a "severe threat".

A five-year academic plan drafted by the university administration calls for the closure of the masters and PhD programmes in physics and astronomy, and for the university's Wyoming Infrared Observatory (WIRO) to be sold or leased. The observatory's 2.3-metre infrared telescope is the third largest in the country.

The plan recommends that consideration be given to scrapping the bachelors degree programme and offering physics courses only as a 'service function' to other departments, such as engineering and geology. The philosophy department and some language programmes are also targeted for closure.

According to the plan, "numerous external reviews, including the most recent WIRO review, have questioned the [economic] sustainability of the physics/astrophysics program". The number of physics majors has dropped from 40 in 1993 to 25 today, and the



number of faculty members is down from 17 two years ago to 13, with only five expected to remain after September.

A recent outside review of WIRO identified deficiencies including infighting among the astronomy faculty members. Although acknowledging past problems, Johnson says the move to eradicate physics reflects a "new climate in academia" focused on cost-benefit calculations. "It flabbergasts me that something at the core of science" could be dropped from the university, he says.

The WIRO review pointed out management flaws and a need for more modern instruments, but it concluded that the observatory is an "important and competitive

facility... which has the potential to excel in certain key scientific niches".

Johnson says he would like to see the observatory run by a consortium of universities. The observatory is negotiating with Denver University in Colorado and other institutions.

Harley Thronson, who heads NASA's Origins programme to study evolution in the Universe, was a member of the faculty at Wyoming until the mid-1990s. He says the department, although small, has produced independent-minded scientists who have risen to "some of the most senior positions in the astronomical community".

James Madison University in Harrisonburg, Virginia, drew national attention in 1995 when it proposed — but later shelved — a plan to shut its physics department. Patrick Mulvey, who tracks data on university degree programmes for the American Institute of Physics in Washington, says he has noticed no trend in the number of physics departments closing down. In a typical year, he says, one or two colleges eliminate or scale down their physics programmes while another one or two expand.

But if the plan is approved, Wyoming would become the only state without a college offering physics degrees. That would make it a scientific backwater, says Johnson. Thronson agrees: Wyoming would lose its influence in astronomy at a time when the field is enjoying a golden era, he says.

Tony Reichhardt

Geological problems drive up cost of nuclear waste dump in Japan

[TOKYO] The Japanese government may have to revise its disposal programme for radioactive waste, following a report from the Ministry of International Trade and Industry (MITI) pointing out the expense of building an underground repository for high-level nuclear waste.

The interim report, released last week by MITI's committee on atomic energy, claims the cost of underground disposal to be eight times more expensive than in France and 13 times more than in Britain. MITI attributes the high estimated cost to extra work required to overcome the problem of high groundwater levels in Japan.

The cost of deep underground disposal in Japan is estimated to be ¥2.7 million (US\$22,460) per cubic metre. According to the report, disposing of 15,000 tonnes of nuclear waste — an average amount produced from dismantling a light-water reactor — could cost ¥19.2 billion.

Given the geological constraints, the construction of the repository would require detailed planning. This would include protective measures against earthquakes to minimize the risk of

radioactive material leaching into groundwater.

With a nuclear waste reprocessing facility currently under construction in Rokkasho village, Aomori Prefecture, the government is planning to dispose of nuclear waste in a deep underground repository which it hopes to build between 2030 and 2040 (see *Nature* 379, 478; 1996).

Japan Nuclear Fuel Cycle Development Organization (JNC), formerly the Power Nuclear Fuel Development Corporation, which plays a central role in investigating the feasibility of deep disposal, is expected to prepare a progress report on nuclear waste disposal by the end of this year.

Based on this report, the government will evaluate the technological reliability of deep disposal by next year, when it plans to establish an agency to carry out the project.

JNC is investigating the 'multi-barrier' system, consisting of engineered barriers (such as vitrified waste and buffer material) and the natural barrier provided by geological formations, as the main model for the planned repository.

But the location of the underground

facility remains undecided. And the expense of the programme is likely to become an additional problem, particularly with large cost overruns and delays in the construction of the reprocessing facility.

While MITI is exploring the possibility of extending the operating period of existing nuclear reactors, JNC is planning a more economical model for the repository, for example by relaxing some of the regulations governing its construction.

Meanwhile, the Atomic Energy Commission, which advises the prime minister, has set up a committee to study ways of reducing the production of high-level nuclear waste.

Three research institutes, including JNC and Japan Atomic Energy Research Institute (JAERI), are developing a technology to stabilize high-level radioactive waste by using fast-breeder reactors and high-energy accelerators.

According to JAERI, such technology could reduce the production of high-level waste by up to 30 per cent, opening up the possibility of keeping such waste "above the ground".

Asako Saegusa

Cosmic-ray observatory in Argentina gets go-ahead

[MUNICH] Site work for what will be the world's largest cosmic-ray observatory is about to begin in Argentina, after a consortium of 53 institutes in 19 countries signed an agreement last month.

The Pierre Auger Southern Observatory, in Mendoza province, will study ultra-high-energy cosmic particles whose origin is unknown and which are so rare that less than a dozen have so far been detected. They generate showers of secondary particles when they hit the Earth's atmosphere.

The US\$50 million observatory will comprise 1,600 surface detectors placed at 1.5-kilometre intervals over an area of 3,000 square kilometres, complemented by four optical telescopes. The detectors — water tanks equipped with photomultipliers — will measure the so-called Cherenkov radiation generated in the water when the secondary particles hit at Earth level. The telescopes will detect the luminous traces in the sky created by particle showers.

The combined data will allow the direction, energy and mass of the primary particles to be determined much more accurately than at smaller cosmic-ray observatories, such as those in Akeno, Japan, and Utah, USA.

The largest single contributor to the Pierre Auger project is Argentina. The central and Mendoza governments have between them put up US\$15 million, which will be used primarily to prepare the infrastructure of the site, an elevated desert plain just east of the Andes.

Roads and a research station will be built in the province's capital, Malargüe. Data from components scattered around the region will be sent to the research station using mobile-telephone technology.

Brazil, France, Mexico, Britain and the United States will work with Argentina on

the water-tank array, while Germany and Italy will provide the telescope technologies.

The observatory should be complete by 2003. According to a spokesman, Alberto Etchegoyen, of the National Atomic Energy Commission's Tandor Laboratory in Buenos Aires, data may be collected as soon as 2001, when many of the detectors will be in place.

Although low-energy cosmic-ray particles with energies of a few million electron volts are plentiful — one particle strikes every square metre of the earth every second — ultra-high-energy particles of over 10^{20} electron volts, which interest scientists in the Pierre Auger collaboration, strike one square kilometre every century. "It is therefore obviously important to cover as much of the sky as possible," says Hans Blümer, a spokesman and head of the Institute for Nuclear Physics at the Forschungszentrum Karlsruhe.

The collaboration decided in 1996 to build a 'twin' Northern Observatory in Utah, of the same size, as soon as funding is found for the project.

For Blümer, the project is a triumph of "grass-roots science". As a large-scale project organized by a consortium of small institutes, he says, "no single very large institute is in charge". He praises the organizational skills of the project's pioneers, Nobel laureate James Cronin, head of the Enrico Fermi Institute at the University of Chicago, and Alan Watson, professor of astrophysics at the University of Leeds in Britain, who first proposed the project eight years ago.

Etchegoyen says the observatory will provide a tremendous boost for young Argentinian scientists. He believes it will encourage funding agencies to create more fellowships for young astrophysicists, which is important "because the physics community here is growing old".

Alison Abbott

Court suspends pioneering gene deal in Yellowstone

[WASHINGTON] A pioneering but controversial agreement that would have allowed Yellowstone National Park to collaborate with a Californian biodiversity company has been suspended by a US court. The plan to exploit the commercial potential of the park's biota was blocked for failing to provide a full environmental impact statement.

Judge Royce Lamberth of the US District Court for Washington ruled last week that the National Park Service, which runs the park, should subject its agreement with Diversa of San Diego to a formal environmental assessment.

The assessment, which will take place under the terms of the National Environmental Protection Act, will require the park service to compare the environmental impact of the agreement with alternative policies.

Yellowstone struck the deal — announced by US vice-president Al Gore in 1997 — to ensure that it will share in any bonanza resulting from the commercial use of genetic material from its plants and microbes.

The park's geothermal springs were the original home of the enzyme *Taq* polymerase, which is used worldwide in polymerase chain reaction (PCR) equipment. Hoffman-La Roche holds patents for its use and the national park stands to gain nothing from its huge commercial success.

But critics say that the deal is not in keeping with the mission of the national parks, and last year sued the park service to halt it (see *Nature* 392, 117; 1998).

Judge Lamberth ruled that the agreement amounted to a significant change of US government policy on bioprospecting on land owned by the federal government, and as such should be subject to public review.

The judge did not rule on other areas of contention, including whether the royalty payments should be published. Immediately after the ruling, the park service said it would suspend the agreement until the environmental assessment was completed.

Jay Short, chief executive of the San Diego-based company, responded to the judgement by announcing that Diversa, while complying with it, would voluntarily pay royalties to Yellowstone from any products which it derives from the bioprospecting it has already undertaken at the national park, pending the environmental assessment.

The small environmental groups that have attacked the agreement welcomed the judgement. "It effectively prevents the exploitation of our national parks solely for commercial gain," says Joseph Mendelson, legal director of the Center for Technology Assessment in Washington. **Colin Macilwain**

Canadian scientists win over rare species

[MONTREAL] Canada's federal government has announced that eight further scientists are being added as voting members to the committee that identifies which species are assigned to the country's endangered list.

The move comes in the wake of a letter written by more than 600 scientists to the prime minister, Jean Chrétien, complaining that political considerations prevented scientific principles from being followed in determining which species should be listed as endangered (see *Nature* 398, 9; 1999). It said the government had abandoned habitat protection as an element of the legislation.

"In effect, [they are] restoring the vote

that was taken away. That's great news," says Jamie Smith, of the University of British Columbia's Centre for Biodiversity Research, who helped draft the letter.

The eight scientists whose votes are restored will chair specialist groups on the Committee on the Status of Endangered Wildlife in Canada, which includes representatives from federal, provincial, territorial and private agencies along with independent experts.

Christine Stewart, Canada's environment minister, said she believes that the change to the committee will "ensure its continued scientific integrity". **David Spurgeon**

US concern grows over secrecy clauses

[BOSTON] A US physician told a conference on secrecy in science this week that he is to lose his job because of claims that he broke a 'non-disclosure' agreement over findings about a potentially fatal occupational lung disease among a group of textile workers.

David Kern, from Brown University, Rhode Island, was speaking at a meeting on Monday (29 March) at Massachusetts Institute of Technology (MIT). Representatives from universities, industry and government had met to address an issue of growing concern as increasing numbers of university researchers form links with private industry.

Two years ago, Kern was told by a Rhode Island textile manufacturer for whom he had done consultancy work that he would be sued if he and his colleagues presented their findings about a lung disease they had identified among company workers. The company claimed that Kern was prohibited from publicizing his research owing to a non-disclosure agreement he had signed.

Kern maintains that the agreement only pertained to a one-hour visit to the plant that he and medical students had made 15 months before the research began. Kern's research contract with the company stated that "we will report to the scientific and public health communities as we deem fit".

But, according to Kern, Brown University and the Memorial Hospital of Rhode Island, at which he works, sided with the company, and his contract with both the hospital and university will be terminated in three months.

The conference, sponsored by MIT and the American Association for the Advancement of Science, examined how university-industry agreements can affect the conduct and scope of scientific research, and the dissemination of findings.

"Secrecy is not always bad," said John Deutch, MIT professor and former director of the Central Intelligence Agency. "Indeed, the privacy of individuals should be protected. But secrecy is a major threat to science and is antithetical to the purposes of universities."

Of particular concern to conference participants were university-industry relationships that could lead to the suppression of research results, delay the publication of data, or otherwise restrict the sharing of information and research tools to allow a company or individuals to preserve a competitive edge.

"Many people are increasingly fearful that the 'non-disclosure agreements' imposed on scientists by industry are fundamentally distorting the way science is done," said Boyce Rensberger, formerly a reporter at *The New York Times* and *Washington Post*, and now director of the Knight Science Journalism Fellowship programme at MIT.

Rensberger believes that non-disclosure

agreements have had a chilling effect on science. "I used to argue that we could count on academic scientists to tell us the truth because they were independent and honest, but nowadays I'm not so sure," he said.

"A fair number of them are subject to gag rules, often saying, 'Sorry I can't tell you; that's proprietary'. When you interview a scientist, you now have to ask: 'Have you signed a non-disclosure agreement that prevents you from speaking freely?' Then you have to decide whether you can trust the answer."

Whether because of non-disclosure agreements or the lure of financial gain, significant numbers of scientists are withholding data or delaying its publication. Alan Goldhammer of the Biotechnology Industry Organization said the problem cannot be blamed entirely on greedy corporations. "Part of the fault lies with the academics who

think, 'By God, I'm going to be rich.'" But Howard Schachman, of the University of California, Berkeley, argued that money has had an unquestionably disruptive influence.

"Why do you need a licence to obtain a reagent from a colleague at another university?" he asked. "The increasing commercialization of research has blurred the line between research tools [such as the polymerase chain reaction] and products."

Although no firm solutions emerged from the conference, the general consensus was that the fruits of research should be shared as quickly as possible to benefit both science and society.

Lita Nelsen, director of MIT's Technology Licensing Office, said academic institutions should establish policies to protect faculty members who lack the power to set appropriate terms with companies. **Steve Nadis**

UK research councils look to future needs

[LONDON] The heads of Britain's main research councils agreed on Tuesday (30 March) to 'steer' a group developing a long-term technology strategy for the whole science base as part of the government's Foresight initiative.

The overall aim of the initiative, launched in 1993, is to stimulate the thinking of scientists and industrialists about how science could help to deliver the products and services of the future, focusing on technologies that create wealth and improve the quality of life.

So far, some areas of science — such as particle physics — have played virtually no role in the Foresight process. Following consultation with the scientific community, however, the government's Office of Science and Technology (OST) recently created 'associate programmes' to allow special interest groups, such as learned societies and research councils, to play a wider role.

A working group of research council officials has already formed a partnership to develop proposals for an associate programme known as the Long-term Technology Review. This will examine the technologies required in five or more years' time to maintain a strong research base in the United Kingdom.

Over the next six months, each research council will analyse its future requirements, based on its science objectives, and produce a statement on its technology needs. These will be drawn into a common report identifying the needs of the science base in specific technical areas, such as detector systems, electronics, materials, biochips and information technology.

This report will allow other Foresight



Halliday: exciting collaboration.

panels to take into account the needs of the science base, and enable common areas of interest to be identified across the science community. An open consultation will also enable researchers to contribute ideas on future technology requirements.

The Particle Physics and Astronomy Research Council (PPARC) has been invited by the OST

to lead the associate programme on the technology needs of the science base on the strength of its work last year on a long-term technology plan for particle physics and astronomy.

Peter Fletcher, PPARC's head of industrial and international liaison, says the first Foresight exercise "did not give us the opportunity to offer solutions to other people's technology problems, although our research requirements are so demanding we are at the cutting edge of technology".

Ian Halliday, PPARC's chief executive, says that the current exercise is "an opportunity for the academic and university community to get into print their requirements. This is a serious mechanism for collaboration between different research councils. We are very excited about it."

The technology review will run for two years and Fletcher hopes that by the end of the process it will have been possible to "create a new understanding of common technology interests between the science base and industry". **Natasha Loder**

Biotech industry seeks 'honest brokers'

[LYONS] The momentum for the development of biotechnology has been seriously set back in Europe by the consumer backlash against genetically modified crops, leading industrialists last weekend. Public confidence, they acknowledged, is a prerequisite if the wider fruits of biotechnology are to be achieved.

This was one take-home message from a four-day meeting in the French city of Lyons which brought together hundreds of delegates, including the chief executive officers of major life-science companies such as Monsanto and Novartis, several Nobel laureates, including David Baltimore, president of the California Institute of Technology, and many leading biotechnologists and geneticists.

The meeting, named 'Biovision', was organized by Raymond Barre, mayor of Lyons and former French prime minister, in conjunction with the US and French academies of science and international agencies such as the World Health Organization (WHO) and the Organization for Economic Cooperation and Development (OECD).

Billed as the first of a biannual life-sciences equivalent of Davos — the international meeting of politicians and industrialists held each year in the Swiss city of that name — it is intended to create dialogue between the warring factions in the biotechnology debate.

But the few politicians or public figures present, such as Lionel Jospin, the French prime minister, and Gro-Harlem Brundtland, the director-general of WHO, just flew in, gave their speeches, and flew out. Environmental and consumer groups were outnumbered by pro-biotechnology industrialists and scientists.

The debates, frequently heated, covered controversial areas of biotechnology and were often relevant to developing countries. Florence Wambugu, director of the Kenya-



Double trouble: the biotechnology industry and Greenpeace have different visions of the future.

based International Service for the Acquisition of Agri-Biotech Applications, for example, attacked opposition to gene technology as a northern luxury. "The biggest risk in Africa is doing nothing," she said. "I appreciate ethical concerns, but anything that doesn't help feed our children is unethical."

The meeting was dominated by discussion of agricultural genetics, bioethics and public perception, in particular the backlash in Europe against genetically modified crops. Some delegates, such as Joël de Rosnay, director of strategy at the Cité de Sciences et de l'Industrie in Paris, maintained that this was simply due to the ignorance of the public, and would disappear if means could be found to educate them.

But Sheila Jasanoff, professor of science and public policy at the University of Harvard, described such attitudes as a "misconception". And there was widespread agreement that the recent advertising campaign by Monsanto to 'inform' the public had only served to deepen public suspicion.

Heinz Imhof, president and director-general of Novartis Seeds, said that companies were perhaps asking for trouble in introducing crops first that, although valuable to farmers, provided few benefits to consumers. The belief among European consumers that genetically modified organisms were being imposed had to be "taken account of" he said, endorsing, as did the meeting, the need for a workable system to label such foods.

The main message of the meeting was that the public's perception of biotechnology is the main challenge to be overcome if the industry is to realize its full promise in terms of improving crop yields and boosting their nutritional and other qualities.

To this end, it identified a need for 'honest brokers' trusted by the public. It concluded that much of the greater acceptance of biotechnology in the United States than in Europe is due to greater public confidence in the independence and competence of the regulatory and scientific advisory system. The US public tend to trust agencies such as the Food and Drug Administration, it was argued, whereas this trust was lacking in Europe after the crisis over bovine spongiform encephalopathy (BSE).

Independent scientific advice was key to public trust, said Jospin, adding that he planned to set up a review of biotechnology decision-making in France.

John Pattison, chairman of the UK government's spongiform encephalopathy committee, said he would transmit to government the need for independent and trusted scientific bodies if public trust was to be rebuilt.

The meeting concluded that ethics committees provide a forum for public debate of such issues. But it also recommended the creation of a European Foods Evaluation agency, as well as independent European and international technology assessment bodies. Brundtland added that WHO intended to set up such a unit to guide its policy setting, by producing global epidemiological analyses and analysing health priorities, policy options and the outcomes of health initiatives.

It was concluded that scientists should anticipate and publicize issues, such as cloning, before they become the focus of media attention. The meeting also suggested that international guidelines for biotechnology research should be established that the public would understand. Progress in these directions should be reviewed at Biovision 2001, it said.

Declan Butler

Asians link up to protect plant resources

[NEW DELHI] Seven countries of the South Asian Association for Regional Cooperation (SAARC) reached an agreement last week not to pass on the plant genetic resources that they exchange among themselves to any country outside the association.

The agreement was reached at a four-day meeting of SAARC countries held to evolve joint strategies for dealing with the issues of plant genetic resources and intellectual property rights, and "to preserve and protect the plant genetic resources of the region". They also agreed to cooperate in addressing the emerging biosafety issues related to genetically modified organisms.

The seven countries — Bangladesh,

Bhutan, India, the Maldives, Nepal, Pakistan and Sri Lanka — agreed to exchange plant germplasm for research purposes freely, but only among themselves. Initially, rice, wheat and maize varieties will be exchanged.

The seven countries also agreed to exchange experts as well as information, and to hold an annual meeting to discuss issues raised over plant genetic resources. But an offer by India to use its gene bank, which has a capacity for one million seed types, for storing the germplasm of SAARC countries was turned down. An Indian official said members wanted their scientists trained in India, but preferred to build their own gene banks.

K. S. Jayaraman

Women scientists unite to battle cowboy culture

Developmental biologist Nancy Hopkins conquered the 'Wild West culture' of MIT, but other women are still fighting. After last week's report on sexual prejudice at the institute, she argues the case for 'institutional change'.

Steve Nadis

[BOSTON] Nancy Hopkins is trying to determine the 'development genes' critical for making a vertebrate. A molecular biologist at the Massachusetts Institute of Technology, she is addressing the question by studying tiny zebrafish, creating mutant breeds with retroviruses and finding the defective genes.

Over the next two years, she and 25 co-workers in her laboratory plan to examine 1.2 million zebrafish. It's a huge undertaking costing more than \$2 million a year, and requiring a lot of space. Five rooms in her lab are devoted solely to housing the fish, in some 4,000 tanks.

Yet it has taken a long struggle for her to muster the resources necessary for this experiment. Since joining the MIT faculty in 1973, and especially since becoming a tenured professor in 1982, Hopkins has faced a system stacked against senior women researchers.

In her battles with the establishment, however, she has helped forge a movement that is now benefiting other MIT women scientists. The problem of sexual discrimination at the institute was formally acknowledged in a report presented to the faculty last month. Hopkins chaired the committee that issued the report and drafted the document.

The publication of the report was a milestone, as MIT publicly owned up to a long-standing pattern of gender bias. And it struck a chord: Hopkins has heard from women scientists around the world who describe similar experiences.

It started when Hopkins discovered that the problems she encountered were common to other women scientists on campus.

Her first battle had to do with laboratory space. About six years ago, shortly after switching from virology to developmental biology and starting her work with zebrafish, Hopkins tried to expand her lab by a modest 200 square feet, but encountered numerous obstacles. Male investigators were sometimes getting whole new buildings while she couldn't even get a closet.

One problem, she says, was "the perception that women liked smaller labs". But she desperately needed more space, and her lab

was smaller than that normally given to starting professors. Although the matter was finally resolved months later, she says, the lengthy ordeal had "sapped her strength".

Then in 1994, two male professors took over a popular biology course that she had developed, and turned it into a book and CD-ROM. Not only was she hurt by the lack of collegiality and respect, but she then had to put a lot more time into designing a new course, taking her away from research. "Time is everything in science," she says. "A few months or a few weeks can cost you the Nobel Prize."

In the wake of this incident, she wrote to MIT president Charles Vest, complaining about the environment that prevailed in science at the institute, which she describes as "a Wild West culture where the strong take from the weak".

Paying a high price for success

Hopkins and two female colleagues, who endorsed her letter, then polled other tenured women scientists at the institute. They were surprised to find there were only 15 tenured women in science, compared with 197 men.

"What was so amazing was that each of us thought we were the only ones struggling," says Hopkins. "These were all gifted women, doing incredibly well in their fields, yet paying a high price for it."

A committee headed by Hopkins was formed to investigate the problem. The members unveiled a process that routinely led to the "marginalization" of senior women faculty, who were kept from the highest levels of power in their departments.

Whereas significant resources are devoted to launching the careers of junior faculty, both male and female, senior women did not have equal access to MIT's vast resources, and often did not even know these resources were available.

Hopkins experienced this first-hand in 1995, when she read in a newspaper that Amgen (which now supports her research) was soliciting proposals from MIT researchers. She asked an administrator how to apply, but he merely replied that a committee had not yet been formed. At the same



Nancy Hopkins: battled a masculine culture in which "the strong took from the weak".

time, the administrator was urging a male colleague to apply for the grants.

She calls this a classic example of why the women's group had to form. "I had to spend so much time fighting for things that others — that is to say, men — got as routine," she explains. "Science is hard enough, but this kind of stress makes it much harder."

Although the committee found that women scientists at MIT are paid less than men of equal standing, "this fight has never been about money," Hopkins says. "What is important to us is our ability to do research."

Life has now improved dramatically for the 15 tenured women, with the support of Vest and Robert Birgeneau, the dean of science. Their salaries have increased by an average of 20 per cent, and they no longer have to raise part of their salaries from grants. They have received help in finding space and resources for their research. The institute has vowed to hire more women scientists, to reflect a student body in which there are now nearly as many women as men.

"Before and after are like two different lifetimes," says Hopkins. Yet she's not certain that conditions have changed in general for women researchers. "This endeavour was essentially secret and focused on the problems of 15 individuals," she notes.

The trick is to make the changes "permanent and institutionalized", she adds. "We can't expect to change everyone's heart and minds, so we have to change the system to make it fairer." Harder still will be to change the cultural climate. "To get ahead here, you have to be so aggressive," she says. "But if women are too aggressive they're ostracized and if they're not aggressive enough, they have to do twice the work." □

Chinese-Americans angered by backlash over Los Alamos leak

[WASHINGTON] Chinese-American physicists have asked the president of the American Physical Society (APS) to speak out in support of their contribution to science, to counter what they call "the deterioration of [their] working environment" following allegations that a Taiwanese scientist leaked US atomic secrets. The accused scientist worked at the Los Alamos National Laboratory in New Mexico (see *Nature* 398, 96; 1999).

The Overseas Chinese Physics Association, which represents 450 Chinese-American scientists, wrote to the APS president asking for his help after a meeting of the group before the society's centennial meeting in Atlanta, Georgia, last month. Group members are angered by what they regard as the unfair characterization of the Chinese-American scientific community in the US press.

The APS is already concerned about increasingly harsh State Department controls on visas for foreign scientists visiting the United States. At least two prominent scientists, from China and Russia, were refused entry to attend the

convention of the International Union of Pure and Applied Physicists, which was also held last month in Atlanta.

Europe's Eureka backs fewer research projects

[MUNICH] Eureka, the European organization that promotes collaboration between companies engaged in research and development, has seen its project portfolio drop by 32 per cent over the past five years.

A review of the 15-year-old organization published last week reveals that the number of participants has fallen by 16 per cent. And the average duration of a project is now 27 months, compared to four years in the early part of this decade.

Companies, which were widely consulted in the review, support the organization's 'bottom-up' approach, but want a stronger voice in its future direction, according to the review.

Japan call for tough limit on dioxin emissions

[TOKYO] The Japanese government last week proposed to reduce emissions of dioxins by 90 per cent from 1997 levels over the next four years. A policy proposal released by the

cabinet panel on dioxin emissions calls for tough regulations on waste incineration, the main source of emissions. It also stresses the need for a new standard for the tolerable daily intake (TDI) of the chemical.

The move comes in response to criticisms of the government's failure to set a common TDI standard. The Environment Agency says the intake of dioxins should be limited to 5 picograms per kilogram of body fat per day, but the Ministry of Health and Welfare sets its TDI limit at 10 picograms.

Russia's genome workers adopt ethics guidelines

[MOSCOW] Russian scientists participating in the Human Genome Project have adopted a declaration of ethical principles on human genome research and its linked medical procedures. This emphasizes that the interests of patients, donors and others taking part in an experiment should override those of society or science.

The declaration says Russian participants in the project should consider it obligatory to observe relevant international conventions and declarations, such as those approved by the United Nations Educational, Scientific and Cultural Organization and the World Health Organization. It adds that any

interference in the human genome is allowed only for medical purposes and provided it does not affect the genes of descendants, and does not lead to genetic discrimination.

China's president visits particle physics lab

[MUNICH] The European Laboratory for Particle Physics (CERN) in Geneva last week received visits from the president of China, Jiang Zemin, and the president of Switzerland, Ruth Dreifuss. They were shown around the 'L3', a major CERN experiment which studies the properties of particles produced by the collision of electrons and positrons.

In 1985, when he was mayor of Shanghai, Zemin helped to organize the production at the Shanghai Institute of Ceramics of crystals which form a fundamental part of the L3 experiment. The first formal cooperation agreement between CERN and the Chinese Academy of Sciences was signed in 1988. China is also participating in CERN's next big project, the Large Hadron Collider.

US nuclear waste store starts work at last...

[WASHINGTON] The first permanent nuclear waste repository in the United States, the

Waste Isolation Pilot Plant in southern New Mexico, has finally received its first truck load of transuranic radioactive waste. The 25-year-old deep-store project finally became operational after the failure of last-minute legal challenges by environmental groups. The first shipment was sent in on 26 March from Los Alamos National Laboratory.

The waste to be stored at the pilot plant initially is only slightly radioactive, consisting of contaminated clothing and materials (see *Nature* 393, 199; 1998). The plant is not licensed to take spent nuclear fuel or other highly radioactive waste, which the US government hopes to store at a planned repository at Yucca Mountain, New Mexico.

... as UK geoscientists back cautious approach

[LONDON] Geoscientists must put more effort into raising public awareness of the strengths and weaknesses of geoscience, particularly in predicting the long-term performance of an underground repository for nuclear waste. This is among conclusions issued last week following an earlier two-day meeting organized by the Geological Society of London.

The meeting had been called to assess the potential contribution of geoscientists in assessing options for managing radioactive

waste. The society concludes that underground disposal remains the safest option, in line with a report from the House of Lords Select Committee in Science and Technology, published last week.

But it emphasizes that "geoscientists do not claim to make unqualified predictions of long-term geological evolution". Rather, they describe the most likely range of future scenarios.

US senators seek \$12m to help save coral reefs

[WASHINGTON] Influential US senators are supporting a measure that would authorize \$12 million in funding for coral reef conservation, in response to the extensive and unexplained wave of damage sweeping the world's coral reefs.

John McCain (Republican, Arizona), chair of the Senate Commerce Committee, and Olympia Snowe (Republican, Maine), who chairs its subcommittee for oceans and fisheries, introduced legislation last week that would pay for additional coral reef monitoring, research and conservation.

"Last year, coral reefs suffered the most extensive damage ever recorded," says McCain. "We must provide funds so that we can begin to understand the causes of such damage."

nature

the international weekly journal of science

invites applications for the post of

Chief Editor, Biological Sciences

Do you have what it takes to lead a growing team of talented editors publishing the world's finest biological sciences papers? To take on the most demanding and stimulating editorial role in biological sciences publishing? To be responsible for the highest standards in assessment and review? To represent *Nature* in the biological sciences community – and to ensure that *Nature* serves the interests of that community?

If so, we want to hear from you.

The Chief Editor, Biological Sciences, is a senior appointment, reporting directly to the Editor, backed by a strong commitment to investment and expansion. He or she will play a key part in developing *Nature's* publishing policy. Terms and conditions will be highly competitive, reflecting the critical importance and responsibilities of the role.



The ideal candidate will probably have some editorial experience, but will certainly have intellectual vision and strong leadership qualities. He or she will have significant achievements in biological sciences research; demonstrable interests across a wide range of biological sciences; respect in the scientific community; a capacity for prompt and balanced critical assessment; good communications skills; and imaginative ideas for the future development of *Nature* and their contribution to it.

The job is based in *Nature's* modern London offices but will involve international travel to meetings and laboratories.

Applicants should write in confidence, marked 'Personal', to:

Dr Philip Campbell,
The Editor, *Nature*,
4 Crinan St, London N1 9XW, UK.

They should include a CV, a discussion of relevant experience, and an indication of their specific ambitions for *Nature*. Applications should be sent as soon as possible and no later than 16 April.

Budget shock hits Danish universities

Sir— Natural sciences faculties in Danish universities are threatened with large reductions in permanent staff. All recruitment of young scientists may come to a full stop, including the transfer of assistant professors to tenure positions, and many scientists in 'permanent' positions can expect to be fired. At the University of Copenhagen about 13 per cent of the staff are expected to be cut.

This is the result of the government budget for 1999. The funding of faculties is based on the number of students. But the support per student varies for historical reasons, without reflecting the costs of experimental work in natural science classes. As a result, the support for each student of natural sciences is markedly lower than that of students in agricultural sciences, medical sciences and civil engineering. The natural sciences have been unable to alter this budgetary arrangement, resulting in a build-up of debt during past years.

Although university professors are expected to divide their time equally

between research and teaching, the budget from the state is determined by the number of students, independent of research activity. So a staff reduction of 13 per cent with an unaltered number of students will reduce the time devoted to research by 26 per cent. The faculties have so far protected research and maintained unaltered dimensions of the different science disciplines independent of the fact that some disciplines have many students, while others have comparatively few. Discussions between the Ministry of Research and the faculties have not resolved the inconsistency between budgetary planning and the teaching and research obligations of the staff. Likewise, discussions within faculties have not altered the balance between students and staff in the different disciplines.

The consequences of the cutback are arbitrary and chaotic for individual laboratories irrespective of their teaching load and international research reputation. Small laboratories with research of

outstanding international quality tend to have a higher turnover of staff and more new tenure positions coming up and can, therefore, foresee a relatively greater reduction in staff. Some laboratories risk losing new research directions completely.

Many recent examples from all over the world demonstrate that political systems find it difficult to provide the investment and the long-term perspective required for high-quality research. The political focus is on management principles and short-term initiatives with an associated public interest which can keep the party in power and the minister in office. As a result, in the future university researchers may attempt to strengthen their contacts with the political establishment, become more visible in the media, and increasingly emphasize the short-term benefits of research rather than the need for long-term planning.

Kaj Sand-Jensen

*Freshwater Biological Laboratory,
University of Copenhagen,
Helsingørsgade 51, DK-3400 Hillerød, Denmark*

High priests of science fear for their position

Sir— I strongly disagree with the tone of your editorial entitled "Ill-advised 'freedom' of scientific information", and with its conclusions (*Nature* 397, 455; 1999).

First, the new legislation is not a "ruse". It is a public law of the United States. Neither was the legislation inserted secretly or in an underhand manner. The authors publicly announced the measure. Second, the federal Freedom of Information Act exempts from the proposed rule files (data) that would constitute a clearly unwarranted invasion of privacy.

Third, data obtained under that act would not be sequestered. It may be reinterpreted. However it is difficult to see how reinterpretation could be considered disruptive. Reinterpretation of data is an integral part of the scientific process. Surely you are not saying that once a conclusion concerning a data set is reached by an author, that conclusion is inviolate! Were that the case the number of pages in *Nature* would be significantly reduced. I also question the use of the word "inquisition". One gets the impression that "well-financed special interest groups" conduct inquisitions, whereas academic colleagues conduct research.

Fourth, I do not understand how

dissemination of scientific information can stifle freedom. On the contrary, it is only through unhindered dissemination that freedom of expression and debate can flourish.

I suggest that we have here not a concern about academic freedom. Rather it is a concern among some that their special status as high priests of 'science' would be subject to review by the non-anointed, with the possibility that an alternate, perhaps superior, conclusion could be reached. It is also a matter of ownership of intellectual property — that is, greed.

George Lauer

*Desert Research Institute,
University of Nevada, 755 E. Flamingo Road,
Las Vegas, Nevada 89132-0040, USA*

Harpoons fly in whale wars

Sir— Frank Cipriano and Stephen Palumbi, in their Scientific Correspondence "Genetic tracking of a protected whale" (*Nature* 397, 307–308; 1999), should have stuck to scientific arguments and left the anti-whaling propaganda to non-scientific publications [see their reply below]. The hybrid blue/fin whale they examined is not "protected" as they claimed: this word cannot be applied to whales taken under permit for research or under an aboriginal

exemption, but only to whales protected from commercial whaling.

Taking any species for scientific research is not a loophole to the protection of whales afforded by the commercial whaling moratorium; it is a right and an obligation applied to all International Whaling Commission (IWC) members by article VIII of the International Convention for the Regulation of Whaling.

IWC resolutions for the genetic testing of whale meat have no legal basis for implementation by IWC. Nevertheless, as Cipriano and Palumbi point out, Norway has already begun to develop a genetic database for minke whales taken by its whalers. Japan is developing a genetic database for all marine mammals landed in the country, whether from small cetacean catches, strandings, or as a result of research.

Cipriano and Palumbi's statement "Thus #26 was killed despite a global moratorium on commercial whaling, and evaded a strict system for the international traffic in protected species..." is provocative, given that this killing, as all agree, was legal. Japan does not permit the import of whale meat from countries that are not IWC members. Iceland resigned from the IWC and is no longer eligible to export whale products to Japan. Norway has prohibited exports of whale meat products.

Cipriano and Palumbi also strayed into propaganda with their statement that "more than 1,000 minke whales are killed each year by IWC members... despite the

moratorium". But the moratorium does not apply to whales taken under scientific permit; nor to IWC members that filed objections to it (Norway); nor to whales allocated by the IWC for aboriginal whaling (Greenland, US and Russian Inuits, and so on). The authors also did not mention that 1,000 minke annually are hardly a statistical blip in a population of more than one million. And they did not explain why they concluded that a "genetic monitoring programme is necessary" because 1,000 minke whales were taken legally.

Finally, Cipriano and Palumbi are wrong to characterize legal rights and obligations as "loopholes... large enough for protected whales to slip through". The catch and identification of a blue/fin hybrid whale advanced scientific knowledge of a species interaction that most scientists had not thought possible. Similarly, the Japanese discovery of a pygmy minke whale species and its minke stock identifications in the Antarctic were significant contributions to science. Would Cipriano and Palumbi ban this type of science to close their "loophole"?

As for the Convention on International Trade in Endangered Species (CITES), more countries would support it if it were more scientifically based and less driven by animal-rights lobbyists. That is why most countries are now trying to reform it.

Alan Macnow

*Japan Whaling Association,
321 East 53 Street, New York 10022, USA*

Palumbi and Cipriano reply — In our view, Macnow does not advance the debate about the role of genetics in whaling or deal with the substantial issues that face efforts to manage international marine resources, nor does he question our results. Our data show how genetic techniques can lead to a positive identification of a particular whale sample. Such an identification system provides the ability to trace an individual from ocean to retail market. This pathway is the focus of all international efforts to manage whaling, but it has never been directly observed until now. Instead, typical management efforts use indirect measures of stock levels and total fishery impact.

The points raised by the Japan Whaling Association via Macnow's letter do not address these issues. Rather, the main point seems to be that whales are not protected because mechanisms such as scientific whaling exist to allow countries to ignore IWC restrictions. Yet the mere fact that a permit is required to conduct scientific whaling shows that the IWC intends to protect species that have been overexploited until populations have recovered.

It may be that the association wishes to conclude that no whale is protected once it has been targeted by a scientific whaling operation because such operations have no

clear bounds. The loose regulations on scientific whaling have been the subject of much debate in the IWC, in *Nature* and elsewhere. Despite the provisions for scientific whaling, IWC and CITES regulations are clearly designed to protect whale stocks from overexploitation. Macnow's opinion that no whales are "protected" is ironically close to our own point that current international regulations allow individuals of species in need of protection to be killed and sold legally.

Macnow also does not accurately represent our call for genetic databases. Our market surveys since 1993 have shown many species available in markets, not just the minke whales taken for research. We recommend a genetic monitoring programme precisely because it is impossible to prove that these whales come from illegal sources without a genetic database of legally taken whales.

However, we are very encouraged by the promise of the Japan Whaling Association not to allow the import of whale products from Iceland, especially after Iceland's decision on 10 March to resume commercial whaling.

Finally, Macnow implies that much scientific whaling is done for valid experimental purposes. This is an important but controversial topic, and Macnow does not provide any solid evidence of the value of data gathered. In fact, non-lethal methods of research have developed quickly over the past decade, and have supplanted the need to kill whales in many instances. Genetic samples from skin biopsies, sloughed skin or even faeces now provide a wealth of information about stock structure that in the past has been provided by liver samples from scientific hunts.

Steve Palumbi, Frank Cipriano

Department of Organismic and Evolutionary Biology, Harvard University, BioLabs, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA

System values seniority over innovation

Sir — As a postdoctoral researcher I can identify with many of the feelings expressed in your poll of postdocs (*Nature* 397, 640–641; 1999). My biggest frustration is with the problem of obtaining external funding. I am sick of having to use a professor's name on research proposals to get them funded.

If I want the all-important funding I have to donate my ideas to a more senior researcher. I receive no recognition, and my control of the budget, and therefore the research design and implementation, is reduced. I currently work in Sweden, and

my friends in Britain mirror my concerns.

It seems that funding councils, staffed by senior researchers, impose a system in which seniority overrides innovation and enthusiasm, and young researchers are open to exploitation. Presumably, should I reach those dizzy heights, I will be expected to boost my research profile in the same way. The system blocks my ideas and career development for those of a professor who, incidentally, I have no doubt is in sympathy with my concerns.

I. A. Brown

*Climate Impacts Research Centre,
Björkplan 6a, 981 42 Kiruna, Sweden*

Curb abuses in graduate school

Sir — Carl Djerassi's Commentary makes the most insightful and useful suggestion I have seen on the issue of mentoring in graduate school (*Nature* 397, 291; 1999).

Djerassi calls for the candid evaluation of professors as research administrators and mentors, a proposal that might genuinely improve the lives of graduate students. He has taken a bold step in addressing this issue squarely and honestly.

I am in a graduate programme at Harvard University that has long employed advisory committees, and the difficulties with them are exactly as Djerassi says. Senior members of faculty hold complete sway over these committees, and other committee members contest the opinion of these individuals at their peril. The committee system is in no way independent or objective, nor does it provide the opportunity to discern, let alone to correct, conflicts or deficiencies in the working relationship between professors and graduate students.

It is widely agreed among graduate students here that Jason Altom would have been no better off with a committee than he was without one. Perhaps this is also recognized by faculty members, but Djerassi is the first to acknowledge it in my experience.

Many students will suffer in their careers, or be driven to abandon science altogether, because of difficulties and abuses in graduate school. In addition to the personal suffering that this engenders, it is an enormous waste of scientific talent. Any effort to provide graduate students with fair, reasonable and considerate advice will not only improve their lot but will be positively reflected in the manner in which they conduct themselves throughout their careers and in their relationships with colleagues and students of their own.

Name and address supplied

Phage-lift for game theory

Martin A. Nowak and Karl Sigmund

The prisoner's dilemma is a classic of game theory in which acting for individual advantage is pitted against acting for collective benefit. An example has been identified among clones of a virus that infects bacteria.

A virus is a natural-born cheat that makes its living by exploiting the vital functions of a host cell. Small wonder, then, that viruses also exploit each other. By neatly combining theory and experiments, Turner and Chao (page 441 of this issue¹) have managed to demonstrate that certain phages — viruses that infect bacteria — actually engage in the prisoner's dilemma, that archetypal trap between cooperation and non-cooperation. Evolutionary game theorists will see this paper as a landmark. Indeed, it will be difficult to find players more primitive than the phage $\phi 6$ and its mutant clone $\phi H2$, stubby chunks of RNA that for their replication depend on a bacterial cell, and are therefore the subject of learned discussions as to whether they constitute life or not.

The prisoner's dilemma was devised by game theorists barely 50 years ago. Today, it seems difficult to conceive how moral philosophers, political thinkers or evolutionary biologists could ever have managed without it²⁻⁴. It is not much of a game, to be honest. Two players each have two options, to cooperate or not cooperate (defect). If both cooperate, they receive a reward, R , which is larger than the punishment, P , obtained if both defect. If one defects and the other cooperates, the defector obtains a payoff, T (the temptation), which is greater than R , and the cooperator receives the sucker's payoff S , which is less than P . So $T > R > P > S$. Because it pays more to defect, no matter whether the other cooperates or not, a rational player is bound to defect. Two rational players, therefore, will each end up with payoff P , instead of the reward R . Lowly $\phi 6$ usually does better and achieves the reward, so one might wonder whether rationality is really the gift it is supposed to be.

As the exemplar for the clash between individual advantage and collective benefit, the prisoner's dilemma was originally used to study the concept of rational choice, and to test actual human behaviour. In 1981, in a seminal paper by Axelrod and Hamilton⁵, it was applied to the evolution of cooperation in biological societies. Axelrod and Hamilton used computer simulations to display the emergence of cooperation in artificial populations, and suggested a wealth of biological examples, ranging from hairless primates

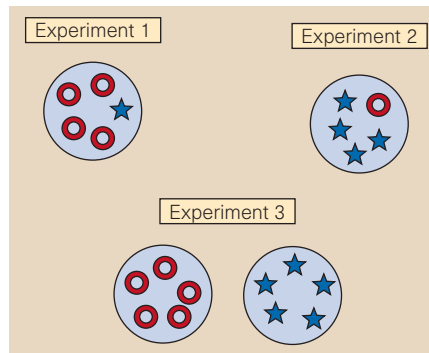


Figure 1 Bacteriophages $\phi 6$ (red) and $\phi H2$ (blue), and the prisoner's dilemma. First, Turner and Chao¹ measured the fitness of $\phi H2$ relative to that of $\phi 6$ in cells that contain mostly $\phi 6$. They thus determined $f(\phi H2, \phi 6)/f(\phi 6, \phi 6) = 1.99$. Second, they measured the same fitness ratio in cells that contain mostly $\phi H2$, and obtained $f(\phi H2, \phi H2)/f(\phi 6, \phi H2) = 1.28$. Third, they mixed cells that contained either $\phi 6$ or $\phi H2$ (but not both) and thereby measured the fitness of $\phi H2$ in cells that contain only $\phi H2$ relative to the fitness of $\phi 6$. They found $f(\phi H2, \phi H2)/f(\phi 6, \phi 6) = 0.83$. Because $f(\phi 6, \phi 6) = 1$ is the point of reference, Turner and Chao obtained the payoff values characteristic of the prisoner's dilemma, as described in the text.

engaged in trench warfare to bacteria living in their guts, in which the principles of the prisoner's dilemma might apply.

In the following years, both computer simulations and study of real-life occurrences of the prisoner's dilemma were expanding areas of research, but they did not grow at an equal pace. It proved much easier to do the simulations, and the empirical evidence lagged sadly behind. The same handful of examples were invoked time and again: vervet monkeys uttering alarm calls, sticklebacks and guppies engaged in predator inspection⁶, vampire bats feeding their hungry fellows. In most cases, the jury is still out on whether these are bona fide instances of the prisoner's dilemma⁷. The underlying problem is the bug-bear of evolutionary game theory: the currency for the payoff values is Darwinian fitness, which is notoriously difficult to measure for monkeys hiding in the bush, bats clustering in cave-roofs, and fish darting in and out of shoals.

With phages, the job becomes doable. The two strategies are embodied in the usual type of $\phi 6$ (the cooperator), and a mutant called $\phi H2$ (the defector) which manufactures fewer of the intracellular products needed for replication of the phages. Turner and Chao¹ measured the relative fitness of the two types in bacterial cultures by means of a genetic marker, cleverly exploiting the fact that the defectors' fitness is greater when they are rare. If the fitness of a $\phi 6$ phage in a $\phi 6$ -infested cell is set equal to 1, then that of a $\phi H2$ phage is almost double ($R = 1$ and $T = 1.99$) (Fig. 1). The fitness of a $\phi H2$ -defector in a $\phi H2$ -infested cell turns out to be $P = 0.83$, and that of a $\phi 6$ -phage in such a cell is $S = 0.65$. This is precisely the rank ordering required for the prisoner's dilemma.

This was by no means a foregone conclusion, even when one strategy is known to be more cooperative than the other. Suppose, for instance, that two cars are caught in a snowdrift⁸. To cooperate means getting out and starting shovelling. If the other driver does this, you can improve your own payoff by defecting (and staying in your heated car). So T is larger than R . But if the other player defects, you are well advised to get out and start shovelling. This is better than spending the night in the car. Hence S is larger than P , in contrast to the prisoner's dilemma.

In this case we end up with the payoff ranking of the so-called chicken game. In an evolutionary setting, defectors will not take over when playing this chicken game. They can invade a population of cooperators, but cooperators can also invade a population of defectors. The outcome is a mixed population. Examples of viral 'chicken' defectors have been known for a while. They quite literally have a defect that means that they cannot reproduce in the absence of complete viruses; in a population consisting entirely of their own kind, their fitness is zero. In this case, S is larger than $P = 0$, and we have a chicken game instead of a prisoner's dilemma.

It seems to us that such a possibility should also be tested carefully in interactions among animals with a cognitive apparatus: in such cases it is likely that a social norm will evolve which determines which of the two players does the cooperating and which does not. For instance, if two lionesses are jointly engaged in territorial defence of their pride, one may consistently be bolder than the other in carrying out defensive duties⁹. If the cost of losing the territory is higher than the risk of an injury, this would be precisely what the theory predicts.

Game theory purists may still quibble that Turner and Chao's matrix does not satisfy the condition $2R > T + S$. This condition is usually added to the definition of the prisoner's dilemma to rule out the possibility that one of the two players cooperates, the other defects, and both then share their total payoff. With phages, this is not to be feared

because it exceeds their capabilities. Another technical objection is that the phages crowded in the bacterial cell are not engaged in a pairwise contest, the usual context of the prisoner's dilemma. Again, this is no serious matter. What counts is that, whatever the ratio of defecting to cooperating phages in a cell, defectors always have an advantage. Necessarily, they will multiply faster, to the detriment of the average fitness.

So why haven't they taken over? Why is the predominant form the good, helpful $\phi 6$ rather than its lazy cousin $\phi H2$? We know little about the ecology of $\phi 6$ (even its natural bacterial host is unknown). But the origin of the defector gives us a clue, for the mutant $\phi H2$ evolves only if the multiplicity of infection is high¹⁰; that is, if there are many phages invading each host cell. In such circumstances, a phage is likely to find itself in a host cell together with phages from another clone, and then it pays to exploit them. In contrast, if the multiplicity of infection is low, the potential suckers are probably closest kin, members of the same clone, and exploitation would be ultimately self-defeating.

It should be noted, however, that the life cycle of phage $\phi 6$ (reproducing in bacteria which eventually burst, and reentering new bacteria) marks it as an ideal candidate for a specific type of group selection¹¹ — which can, more orthodoxly, be viewed as individual-based selection for the ability to build successful groups¹². Is cooperation due to being related or to being in the same boat? Even non-related phages may be better off by cooperating, and producing their full share of intracellular products, if the defective $\phi H2$ has an edge in the competition within the cellular compartment only, and not over the full life cycle. Exploring this issue calls

for further experiments with artificially increased competition based on high multiplicity of infection: these phages may become a testing ground for arguments on kin selection versus group selection, or on the evolution of virulence¹³, just as they have been used for studying sexual recombination¹⁰.

Finally, we recall that the role of compartments in sheltering cooperators from being exploited is crucial to certain hypotheses on the origin of life^{14,15}. Today, phages are arguably the simplest players of the prisoner's dilemma. But it is conceivable that, a few billion years ago, still more primitive molecules were engaged in that game. □

Martin A. Nowak is at the Institute for Advanced Study, Princeton, New Jersey 08540, USA.

e-mail: nowak@ias.edu

Karl Sigmund is at the International Institute for Applied Systems Analysis, Laxenburg, Austria, and the Institute of Mathematics, University of Vienna, Strudlhofgasse 4, A-1090 Vienna, Austria.

e-mail: ksigmund@esi.ac.at

1. Turner, P. E. & Chao, L. *Nature* **398**, 441–443 (1999).
2. Trivers, R. Q. *Rev. Biol.* **46**, 35–57 (1971).
3. Axelrod, R. *The Evolution of Cooperation* (Penguin, London, 1989).
4. Sigmund, K. *Games of Life* (Penguin, London, 1995).
5. Axelrod, R. & Hamilton, W. D. *Science* **211**, 1390–1396 (1981).
6. Milinski, M. *Nature* **325**, 433–435 (1987).
7. Dugatkin, L. A. *Cooperation Among Animals: An Evolutionary Perspective* (Oxford Univ. Press, 1997).
8. Sugden, R. *The Economics of Rights, Cooperation and Welfare* (Blackwell, Oxford, 1986).
9. Heinsohn, R. & Packer, C. *Science* **269**, 1260–1263 (1995).
10. Turner, P. E. & Chao, L. *Genetics* **150**, 523–532 (1998).
11. Wilson, D. S. *Am. Nat.* **149**, 336–351 (1997).
12. Ridley, M. *The Origin of Virtue* (Penguin, London, 1996).
13. Nowak, M. A. & May, R. M. *Proc. R. Soc. Lond. B* **255**, 81–89 (1994).
14. Eigen, M. & Schuster, P. *The Hypercycle: A Principle of Natural Self-Organization* (Springer, Heidelberg, 1979).
15. Maynard Smith, J. & Szathmari, E. *The Origins of Life* (Oxford Univ. Press, 1999).

case in which there is reasonable evidence for beamed (rather than isotropic) emissions of γ -rays^{2,4,9}. And it is the first in which a radio afterglow appeared after just one day, only to disappear the next^{2,3}.

The reason these measurements are so exciting is that, although significant progress has been made in understanding the γ -ray and afterglow emission in terms of the standard fireball-shock model, many skeletons still lurk in the closet. The emission from γ -ray bursts is thought to occur when one or more shock waves form in a relativistically expanding fireball following a cataclysmic explosion. An intense debate continues on whether the initial γ -rays arise in internal shocks within the original outflow of nucleons from the fireball, or in an external shock as the outflow is decelerated by the surrounding medium. Furthermore, an understanding of the 'central engine' of the burst remains elusive, in part because γ -ray bursts emit most of their energy during the first hundred or so seconds, decaying afterwards with a power-law slope. Until now, only γ -ray and some X-ray information has been available for this crucial early period — prompt optical and X-ray measurements are needed to further constrain models of the radiation mechanism.

Another problem is the energy budget. For instance, if GRB990123 were emitting energy isotropically at the measured redshift^{2,6} distance of $z = 1.6$ (corresponding to when the Universe was only 30% of its current age), it would require 4×10^{47} J (equal to the energy contained in two solar rest masses). This would strain to breaking point any model based on a stellar progenitor, such as the collapse of a massive rotating star (a hypernova or collapsar), or a merger involving neutron stars or black holes. The measurement of an immediate optical afterglow¹ and possible evidence for beaming^{2,4,9} provides much needed clues in attempting to address these issues.

The early optical detection of GRB990123 was achieved with a robotic optical camera, called ROTSE, installed on a fast-response swivelling mount. ROTSE consists of four telephoto lenses, which together can rotate and point more quickly (to reach any part of the sky within three seconds) than any large telescope. It responds to burst coordinates arriving by way of the Internet from the γ -ray coordinate network¹⁰, which in seconds transmits worldwide the rough locations (within a five arc-degree error) of all γ -ray bursts detected by NASA's Compton Gamma-Ray Observatory in space. With this instrument, Akerlof *et al.*¹ measured an optical brightness in excess of ninth apparent magnitude — which is too faint to see with the naked eye, but could be detected with a pair of binoculars — peaking 50 seconds after the trigger of GRB990123. In terms of brightness, this energetic burst is in the top

Gamma-ray astronomy

A burst caught in flagrante

Peter Mészáros

Gamma-ray bursts seem to delight in surprising astrophysicists. The most recent example is the event of 23 January 1999, analysed in three papers by Akerlof *et al.*¹, Kulkarni *et al.*² and Galama *et al.*³ in this issue of *Nature*, and in three related papers in *Science*^{4–6}. To appreciate the importance of these observations, we must remember that, for 23 years, γ -ray bursts were just that: brief pulses of γ -rays that pierced, for a fleeting instant, an otherwise pitch-black γ -ray sky. With rare exceptions, they left no trace at any other wavelengths, until in early 1997 the Italian–Dutch satellite Beppo–SAX succeeded in providing accurate X-ray measurements that allowed, after a delay of 4–6 hours, their position to be pinpointed and enabled follow-up measurements at optical and other wavelengths (for

example, ref. 7). This paved the way for the measurement of redshift distances, the identification of candidate host galaxies, and confirmation that they were indeed at cosmological rather than galactic distances (for example, ref. 8). This implies that γ -ray bursts are among the most powerful events in the Universe.

The burst known as GRB990123 is unique in at least four respects. It is the first to be detected optically while it was still emitting γ -rays — 22 seconds after being triggered¹. If GRB990123 emitted energy equally in all directions, then it would have to be the most energetic γ -ray burst detected so far^{2,6}. Alternatively, if the burst beamed its γ -rays in a jet, and an observer was near to the jet axis, then much less energy would be needed to produce the same intense flux. GRB990123 is the first

0.4% of all previously observed γ -ray bursts³, which may help explain why this is the first burst to be detected so early. It also suggests that such bright optical flares are rare events, although the rate of detection may increase with future improvements to ROTSE.

The early optical light curve seen by ROTSE decays more steeply than that of previous optical afterglows detected after a delay of several hours. However, after about ten minutes its decay rate moderates, and appears to join smoothly onto a slower decay rate measured, after several hours and days, by large ground-based telescopes^{2,3} and the Hubble space telescope⁹. The decay rate between ten minutes and three days is indistinguishable from that observed in several previous afterglows. Indeed, an early ninth-magnitude optical flash with a steep initial decay had been predicted^{11,12} by theories involving a reverse component of the shock advancing into the external medium, which propagates back into the relativistic outflow from the γ -ray burst. The idea that the optical brightness arises from an internal shock¹¹ cannot be ruled out, but is less likely¹³, especially if, as appears to be the case, the peak of optical brightness did not correspond with the peak of the γ -rays. The subsequent slower decay agrees with predictions for the forward component of the relativistic shock propagating into the surrounding medium^{11–14}. The radio emission, rising after one day and disappearing the next^{2,3}, is quite unusual. Previous afterglows have become detectable after several days, and have typically remained visible for months. Whereas the radio data are too sparse to allow thorough theoretical modelling, their peculiar character may be

related to the other unusual properties of GRB990123.

The tentative evidence for beamed emission is perhaps the most exciting. It is based on an apparent steepening of the light curve after about three days^{2,4,9}. This is harder to establish than the decay of the two earlier portions of the light curve, because by this time the flux has decreased to a level where the detector noise and the light of the host galaxy become important. If the steepening is real, after correcting for this interference, it is probably due to the transition between early relativistic expansion, when the optical light-cone is narrower than the γ -ray beam, and the late expansion, when the light-cone has become wider than the beam, leading to a drop in the effective flux^{2,13,15}. A rough estimate gives a beam size of 3–5 degrees, which would reduce the total energy output to about 4×10^{45} J. This is about two orders of magnitude less than the binding energy of a few solar rest masses, which, even allowing for substantial inefficiencies, is compatible with currently favoured proposals for the origin of γ -ray bursts (see, for instance, ref. 16) based on a stellar collapse or a compact binary merger.

To make further progress in understanding the roles of internal and external shocks, beaming and the nature of the progenitors will require significant and concerted theoretical and observational efforts. Models should be able to reproduce and predict the details of both the early and delayed radiation at various wavelengths under realistic astrophysical conditions, for example involving a variable energy of ejection of the fireball outflow or an inhomogeneous external medium. The formation of relativistic jets needs to be understood better, as does the

ability of various progenitor candidates to power them. Observational progress will depend on better statistics for the location of transients with respect to the host galaxy, and extensive studies of early optical, X-ray and possibly radio flashes. This requires obtaining spectra within minutes of burst ignition, and following light curves for fractions of hours or longer until large observatories can take over the pursuit. Some of this will be achieved with upgraded versions of ROTSE. But follow-up observations of faint fluxes will require small dedicated satellites such as HETE—a γ -ray/X-ray mission scheduled to be launched next year—or Swift, a proposed γ -ray, X-ray and optical follow-up satellite designed to catch up to 300 γ -ray bursts per year ‘on the fly’.

Peter Mészáros is in the Department of Astronomy and Astrophysics, Pennsylvania State University, 525 Davey Laboratory, University Park, Pennsylvania 16802, USA, and the Institute of Theoretical Physics, University of California, Santa Barbara, California 93106, USA.
e-mail: pmeszaros@astro.psu.edu

1. Akerlof, C. *et al.* *Nature* **398**, 400–402 (1999).
2. Kulkarni, S. *et al.* *Nature* **398**, 389–394 (1999).
3. Galama, T. *et al.* *Nature* **398**, 394–399 (1999).
4. Castro-Tirado, A. J. *et al.* *Science* **283**, 2069–2072 (1999).
5. Hjorth, J. *et al.* *Science* **283**, 2073–2075 (1999).
6. Andersen, M. I. *et al.* *Science* **283**, 2075–2077 (1999).
7. van Paradijs, J. *et al.* *Nature* **386**, 686–689 (1997).
8. Metzger, M. *et al.* *Nature* **387**, 878–880 (1997).
9. Fruchter, A. *et al.* <http://xxx.lanl.gov/abs/astro-ph/9902236>
10. The GRB Coordinates Network <http://gcn.gsfc.nasa.gov/gcn>
11. Mészáros, P. & Rees, M. J. *Astrophys. J.* **476**, 232–237 (1997).
12. Sari, R. & Piran, T. <http://xxx.lanl.gov/abs/astro-ph/9901105>
13. Mészáros, P. & Rees, M. J. <http://xxx.lanl.gov/abs/astro-ph/9902367>
14. Sari, R. & Piran, T. <http://xxx.lanl.gov/abs/astro-ph/9902009>
15. Rhoads, J. *Astrophys. J.* **487**, L1–L4 (1997).
16. Popham, R., Woosley, S. E. & Fryer, C. *Astrophys. J.* (in the press).

Animal behaviour

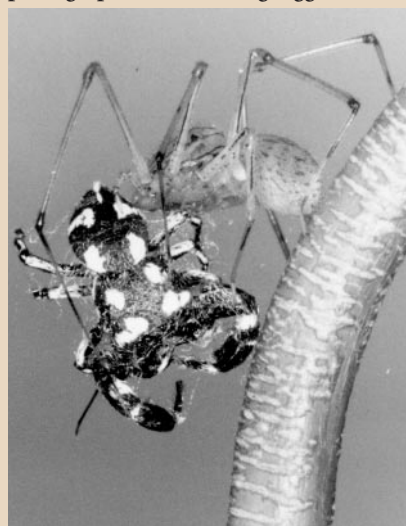
Spit and polish off

Punk rockers were renowned for doing it; some cobras do it; in Jurassic Park, *Dilophosaurus* did it. Now, a Philippino spider has made the unlikely connection between aggressive spitting and diligent motherhood (D. Li *et al.* *J. Zool. Lond.* **247**, 293–310; 1999).

The spider, an as yet unnamed species belonging to the genus *Scytodes*, ambushes prey in its web, or chases it down in the web or on surrounding vegetation. It subdues its victim by spitting a sticky gum over it — taking aim from as much as 60 mm away — before moving in for the kill and wrapping the meal in silk. Large targets are spat at repeatedly.

Li *et al.* found that *Scytodes* feeds mainly on jumping spiders. This is an unusual specialization for a spider — prey sometimes turns predator, necessitating caution and perhaps explaining the

complex hunting behaviour. The photograph shows the long-legged



Scytodes hunting away from the web and plunging its fangs into the jumping spider *Ligurra*, which is covered in a spit net. *Scytodes* is the only genus of spider known to capture prey by spitting.

Female *Scytodes* do not just spit to feed themselves, however. Like many spiders, this species carries its egg sac in its jaws. Li *et al.* show that this behaviour improves the survival of eggs, perhaps by protecting them from ants or desiccation. But unlike most spiders, parental care does not stop there. After hatching, the young remain in their mother's web, and she brings food to them with none of the aggression or cannibalism common in spiders. Truly a noble addition to the pantheon of expectoration, and further proof that complex social interactions emerge everywhere in the animal kingdom when someone looks hard enough. **John Whitfield**

Population biology

Crash tests for real

Robert M. May

In Tom Stoppard's play *Arcadia*¹, which weaves together themes of Byron scholarship, landscape gardening and chaos, along with much human goings-on, one of the characters is doing postgraduate research on the population cycles of grouse. In the play, Valentine gives up; it is all too complicated. Two experimental studies, one on grouse cycles² and the other on vole cycles³, could have changed his mind.

In real life, grouse cycles are broadly acknowledged to be caused by density-dependent regulatory effects, acting with time delays. But there has been great debate about whether these effects derive from the grouse populations' interactions with food supplies or vertebrate predators or parasite infestations or territory size or other factors⁴. Hudson *et al.*² seem to settle these questions, using experimental manipulations to "demonstrate that parasites are the cause of the cyclic fluctuations".

Earlier work by Hudson⁵ led to the suggestion that the marked cyclic fluctuations seen in about three-quarters of the British populations of red grouse (*Lagopus lagopus scoticus*; Fig. 1), which are managed for shooting, are host-parasite cycles, caused by the parasitic nematode *Trichostrongylus tenuis*. In 1988, Hudson and colleagues began a study of six grouse populations, for all of which they predicted — on the basis of host-parasite models^{6,7} — troughs of the population cycle in 1989 and 1993. In two of the populations birds were caught and treated with anthelmintics to eliminate parasites in advance of both of the predicted population crashes. In two populations, no treatments were made. In the remaining two, birds were caught and treated ahead of the predicted 1989 crash, but not ahead of that predicted for 1993.

The results are remarkable. In the two untreated populations, crashes representing a more than thousandfold reduction in grouse numbers (as assessed by counts in the field, and rather differently by shooting-bag records) indeed occurred in 1989 and 1993. In the two twice-treated populations, one remained steady throughout the study (1988–96), and the other showed only a relatively small blip (threefold decrease) in 1993. The two populations treated in 1989 but not in 1993 both showed the predicted major crash in 1993, but in 1989 one remained steady while the other — call it A — showed a roughly tenfold downward fluctuation.

This exception, A, to an otherwise consistent record of anthelmintic treatment converting cycles to steadiness, is illuminat-



Figure 1 Red grouse. Populations of the bird are subject to cyclic variation — driven, it now seems, by a parasitic nematode.

ing. Hudson *et al.*² estimate that, in the four populations that were treated (once or twice), they actually managed to give anthelmintics to somewhere between 15% and 50% of the adult breeding birds. They also used standard mathematical models for host-parasite associations, with parameters assessed from their earlier work^{5,7,8}, to estimate how large a proportion of breeding adults needs to be treated in order to move from cyclic to steady dynamics; this is a variation on the 'what proportion must be vaccinated to eradicate measles' theme. The outcome is a theoretical estimate that more than 20% coverage is needed to eradicate cycles in this system. Hudson *et al.* suspect that the single treatment in the population A, which showed marked reduction in the amplitude but not the disappearance of the 1989 cycle, may have had the lowest proportion of birds treated, at around 15%.

The other experimental study that might have changed Valentine's life is by Korpimäki and Norrdahl³, and deals with small rodents and their predators in Finland. Korpimäki and Norrdahl studied populations of voles (field voles *Microtus agrestis* and *M. rossiaemeridionalis*, and bank voles *Clethrionomys glareolus*) which show pronounced three-year cycles in abundance. Food shortage, predators and other factors all have advocates as the underlying cause of the cycles. Korpimäki and Norrdahl chose six different areas, each of around 2–3 km², in which they reduced the population density of the voles' weasel/stoat and avian predators ahead of the predicted crash phases of

1992 and 1995. Specifically, they reduced numbers of the 'least' weasel (*Mustela nivalis*) and stoat (*M. erminea*) by live trapping and removal from the experimental areas. For the Eurasian kestrel (*Falco tinnunculus*) and Tengmalm's owl (*Aegolius funereus*), they removed all of the nest sites they could find before the breeding season began.

In 1992 Korpimäki and Norrdahl were more successful in removing least weasels than stoats, and they did not attempt to reduce the numbers of birds of prey. In 1995, however, learning from their earlier experience, they believe that they achieved an effective reduction of all these predators. In both years, there was no difference in the initial abundance of predators and the pooled abundance of small rodents between their six manipulated areas and control areas elsewhere. In 1992, with effective removal of least weasels only, the cycle persisted and the voles crashed, as also happened in all control areas in both 1992 and 1995. But, strikingly, the pooled prey populations remained roughly steady in the manipulated areas in 1995, when all vertebrate predators were reduced.

Korpimäki and Norrdahl also refer to an earlier version⁹ of the new work, in which only avian predatory numbers were reduced. Here again they found the voles crashing at the trough of the cycle in both manipulated and control areas. This ensemble of experimental results leads them to conclude that their particular three-year vole cycle is indeed a predator-prey one, but that all predators must be taken into account. Removal of any one subset of predator species — least weasels only or raptors only — may, they argue, leave other predators to continue to drive the cycle.

The Finnish vole study has interesting parallels to that on the Scottish grouse in involving large-scale field manipulation to illuminate a basic ecological question. It would have been even nicer had Korpimäki and Norrdahl attempted a model-based estimate of how much predator reduction was needed to move from cycles to steadiness, along the lines of that done by Hudson *et al.*². Korpimäki and Norrdahl do estimate that least weasels account for very roughly 40% of all predation on their voles, but they unfortunately have no theoretical estimate of the 'cycle threshold' (corresponding to Hudson and colleagues' estimate of 20% parasite reduction) to put alongside this number. I admit, however, that there would be formidable difficulties in parameterizing a multi-predator model for Korpimäki and Norrdahl's system.

Given the long-standing interest in population cycles, the reader may well ask why these kinds of field manipulations were not done years ago. I have no easy answer. I could point out that much of the empirical

CORBIS/ROGER WILMHURST

NATURE

100 YEARS AGO

A deputation of representatives from the Decimal Association, chambers of commerce, educational institutions, and trade unions, waited upon Mr. Ritchie on Wednesday, March 22, at the House of Commons to urge upon the Government the compulsory adoption of the metric system of weights and measures on January 1, 1901. Several of the delegates described the advantages which the metric system possesses, reference being made to the great waste of time involved in teaching our complicated system of arithmetic, and the loss of trade resulting from the use of a system not understood by other nations. In reply, Mr. Ritchie expressed himself in agreement with the arguments in favour of the metric system, but stated that his own view, and that of his colleagues, was that chaos and confusion would be created by the compulsory adoption of the metric system in this country, and it would be practically impossible to carry out a compulsory law on the subject. They had not only passed a law two years ago to make the metric system legal, but they had also added to their Board of Trade standards the standards for the metric system, and only seventeen of the whole of the local authorities in the country had come to verify their standards.

From *Nature* 30 March 1899.

50 YEARS AGO

'Earthquake Willis', one of the most colourful and widely known figures in American geology, has passed away. ... For him the theory of continental drift was "a fairy tale, ein Märchen"; no need to invoke it to answer the demands of palaeontology when temporary isthmian links might do the same with less affront to his concepts of earth-mechanics. For him, the propulsive force in earth-movements was volume-change at depth, brought on by the rise of 'asthenoliths', blisters of molten rock, developed by decay of radioactive substances scattered irregularly beneath the crust. The earth is growing hotter; the rising blisters expand upward, and their covering rocks, becoming foliated by recrystallization, spread sideways. Discoidal surfaces of shear form in the lithosphere; the disks rise and fall. Metamorphism is no less the cause than the result of orogeny.

From *Nature* 2 April 1949.

and theoretical work on host-parasite dynamics is relatively new; this explains why the work by Hudson *et al.* is so recent, but not the vole cycle study. Part of the explanation is surely that such large-scale experimental studies, working with complicated systems in the field, are very difficult. But I suspect that part of the reason is that these kinds of studies are necessarily a bit messy and imprecise, lacking the crisply controlled accuracy of a study on a 10-m² plot which too often seems more glibly 'scientific' to a funding agency. There have indeed been many such studies of population cycles in tiny enclosures, and they have not taught us a lot. I cheerfully see studies such as those described here as part of a growing recognition that important ecological questions simply have to be addressed

on the right scale — which often means an uncomfortably large scale — even if that means accepting a certain degree of imprecision. □

Robert M. May is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

e-mail: robert.may@zoo.ox.ac.uk

1. Stoppard, T. *Arcadia* (Faber & Faber, London, 1993).
2. Hudson, P. J., Dobson, A. P. & Newborn, D. *Science* **282**, 2256–2258 (1998).
3. Korpiimäki, E. & Norrdahl, K. *Ecology* **79**, 2448–2455 (1998).
4. Krebs, J. R. & May, R. M. *Nature* **343**, 310–311 (1990).
5. Hudson, P. J. *Grouse in Space and Time* (Game Conservancy Trust, Fordingbridge, UK, 1992).
6. May, R. M. & Anderson, R. M. *J. Anim. Ecol.* **47**, 249–267 (1978).
7. Dobson, A. P. & Hudson, P. J. *J. Anim. Ecol.* **61**, 487–498 (1992).
8. Hudson, P. J., Newborn, D. & Dobson, A. P. *J. Anim. Ecol.* **61**, 477–486 (1992).
9. Norrdahl, K. & Korpiimäki, E. *Oecologia* **103**, 241–248 (1995).

Semiconductor physics

Phonons rewrite structural recipes

Mark S. T. Bukowinski

The well-developed physics of semiconductors has given scientists fast computers and experimental equipment capable of measurements with high resolution. So it is fitting that, by applying these very tools, physicists have discovered that the domain of semiconductor crystal structures, and the processes responsible for them, is far richer than indicated by the models upon which the advances were built. First, there were new high-pressure angle-dispersive X-ray measurements, which showed that some of the previously predicted and apparently observed structures did not exist¹. Now, writing in *Physical Review Letters*, Ozoliņš and Zunger² show that calculations from first principles may give the wrong structures when such calculations are only based upon the 'usual suspects'. The authors used fast computers to incorporate the effects of phonons — which are quanta of crystal lattice vibrations — into the theory, and discovered that the old standby structures do not stand up well to atomic vibrations.

The prototypical semiconductor is part of the family of $A^N B^{8-N}$ compounds, where A and B are elements in the N^{th} and $(8-N)^{\text{th}}$ main groups of the periodic table. For example, gallium arsenide (GaAs) and indium phosphide are $A^{\text{III}} B^{\text{V}}$ compounds. Observations and simple theoretical models based on concepts of ionic size, as well as the relative electronegativity (power of the atoms to attract electrons) and hence the degree of covalency or ionicity of the chemical bonds, yield three dominant structure types for these compounds³. Those in which the energy benefits of forming four, fully saturated covalent bonds outweigh the competing need for efficient atomic packing, adopt tetrahedrally coordinated structures. These

'covalent' structures include the diamond lattice of elemental semiconductors such as silicon and germanium, and the zinc blende or wurtzite structures of many $A^{\text{III}} B^{\text{V}}$ and $A^{\text{II}} B^{\text{VI}}$ compounds. Those 'ionic' compounds with ionicities that exceed about 80% tend to adopt the octahedral coordination of the sodium chloride (NaCl) structure. The semimetallic tin adopts a third 'metallic' structure, β -tin. This is also octahedrally coordinated, although the four equatorial bonds are somewhat distorted towards tetrahedral corners, suggesting a significant covalent influence. Pseudopotential calculations from first principles, which involve solutions of Schrödinger's wave

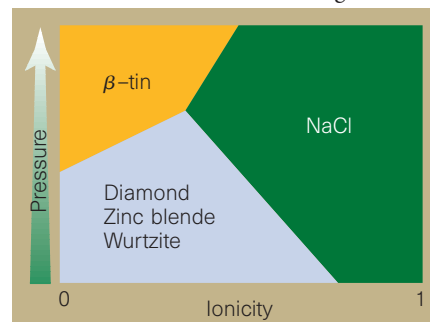


Figure 1 Schematic phase diagram showing how the structures of most of the $A^N B^{8-N}$ octet semiconductors were found to depend on bond ionicity and pressure. (The pressure scale is of the order of 10 GPa.) Detailed structural information, made possible by the combination of synchrotron X-rays with high-pressure diamond-anvil devices, revealed that only silicon and germanium adopt the β -tin structure at high pressure, and that many of the purported NaCl phases do not exist. Observed structures have lower symmetry, often corresponding to distortions of the NaCl or β -tin structures.

equation without any empirically adjustable parameters (only carefully developed approximations of a 'mean field' type are allowed to convert the intractable many-electron problem to a single-electron problem), have largely supported the three-structure model⁴.

This simple picture appeared to remain intact when some of the compounds were subjected to high pressures. Although pressure lowers ionicity, the NaCl structure offers the advantage of higher coordination of atoms and hence more efficient atomic packing, as does the β -tin structure. So a simple phase diagram was born (Fig. 1). Again, electronic structure calculations from first principles were consistent with observations and earlier model predictions⁴. Highly covalent compounds were found to transform directly to the β -tin structure, whereas more ionic ones encountered the NaCl structure as an intermediate phase.

Hints that our traditional understanding might be incomplete emerged in the late 1970s. The semiconductor GaAs was found to transform at a pressure of 17 gigapascals (GPa) into an orthorhombic phase⁵, rather than the expected β -tin. (Atmospheric pressure is approximately equal to 10^5 Pa.) When forced to examine other possible structures, theorists found one⁶ with lower optimal energy than those of the NaCl and β -tin. It is now known that this structure, which can be derived from NaCl by slight displacements of alternate NaCl lattice planes, has a symmetry described by the $Cmcm$ space group.

Such distortions were not included in earlier computations, because the slower computers used then did not encourage lengthy and memory-hungry calculations that minimize energy with respect to multiple degrees of freedom. So theoreticians would minimize the energy of the three prototypical structure types (the usual suspects) within the restricted space of those crystal parameters that would not change the crystal structure, such as the lattice constant of NaCl. The results were satisfying and encouraging, as they accorded with the canonical phase diagram.

As the saying goes, you get what you pay for. Energy optimizations in a restricted parameter space can only yield structures that are consistent with those restrictions. Figure 2 illustrates the potential penalties; structures with different symmetry will be missed. As bad luck would have it, the X-ray spectrum of a slightly distorted structure differs from the undistorted one by small splittings of certain spectral lines. Until recently such splittings were not resolvable. It is ironic that, despite the fact that distortions of this type are common among known solids⁷, it is limitations in semiconductor technology (among others) that, by restricting techniques in computational and experimental semiconductor physics, caused both

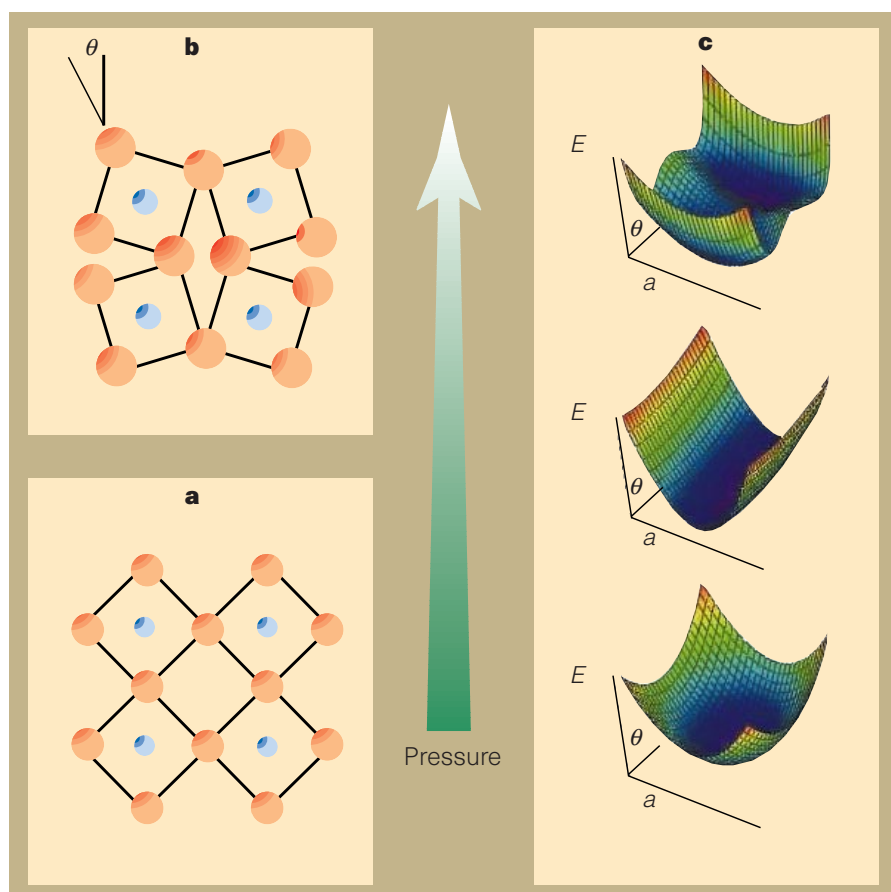


Figure 2 Hypothetical two-dimensional crystal structure and associated energy surfaces.

a, Equilibrium low-pressure structure has square symmetry. b, At high pressure the structure distorts by way of linked rotations, through angle θ , of the square coordination 'polyhedra'. c, The energy surface as a function of the lattice constant a , which is the side length of the unit cell consisting of four polyhedra, and θ . At zero pressure the surface has a minimum at $\theta = 0$. Optimization with respect to a alone yields the correct structure. At intermediate pressures the structure is still square, but the energy depends weakly on θ . At high pressures, restricted optimizations still yield a square lattice; but because the $\theta = 0$ extremum has now become an unstable saddle point (maximum with respect to θ), the lattice distorts in response to any perturbation in θ , such as would be created by a lattice vibration. The squared frequency of the corresponding phonons decreases with pressure because the restoring force (which is proportional to the curvature of the energy along θ) diminishes until, at some critical pressure, the force disappears altogether, and atomic displacements proceed until one of the flanking stable minima is found. The new structure may reflect additional distortions as in b, where the structure shrinks along the horizontal direction, giving two distinct lattice constants.

to see the same limited structural menagerie.

And yet the results reported by Ozoliņš and Zunger² capitalize on recent computational advances. By computing phonon spectra from first principles for several semiconductors, the authors suggest that the NaCl and β -tin structures have intrinsic dynamical instabilities. In NaCl there appears to be a transverse acoustic phonon whose frequency decreases with pressure. So, although the NaCl structure of gallium phosphide is found to be energetically stable relative to the zinc blende structure at pressures exceeding 1.68 GPa, by then the NaCl structure is dynamically unstable. Interestingly, the β -tin instabilities are caused by soft phonons whose frequencies increase with pressure, suggesting that this structure might become stable at higher pressures. The

net outcome is a structurally more diverse family of semiconductors.

Time will tell whether the theoretical details are accurate. There is good correspondence between the latest experimental structures and the results reported by Ozoliņš and Zunger. Whatever the outcome, the new theoretical and computational capabilities greatly improve the quantum-mechanical structural microscope. This will benefit all those who search for new materials or who try to divine the mineral composition of planetary interiors, as well as enhancing our understanding of everyday materials such as semiconductors. □

Mark S. T. Bukowinski is in the Department of Geology and Geophysics, University of California, Berkeley, California 94720, USA.
e-mail: markb@socrates.berkeley.edu

1. McMahon, M. I. & Nemes, R. J. *Phys. Status Solidi* **198**, 389–402 (1996).
2. Ozols, V. & Zanger, A. *Phys. Rev. Lett.* **82**, 767–770 (1999).
3. Phillips, J. C. in *Bonds and Bands in Semiconductors* (eds Alper, A. M., Margrave, J. L. & Nowick, A. S.) (Academic, New York, 1973).
4. Chelikowsky, J. R. *Phys. Rev. B* **34**, 5295–5304 (1986).
5. Yu, S. C., Spain, I. L. & Skelton, E. F. *Solid State Commun.* **25**, 49–51 (1978).
6. Zhang, S. B. & Cohen, M. L. *Phys. Rev. B* **39**, 1450–1452 (1989).
7. Rao, C. N. R. & Gopalakrishnan, J. in *New Directions in Solid State Chemistry* 168–177 (Cambridge Univ. Press, 1997).

Developmental biology

A unity of opposites

Ethan Bier

During development, cells secrete factors — referred to as morphogens — that influence the fates of their neighbours. These morphogens are distributed across groups of cells in a graded fashion, eliciting distinct responses depending on their concentration. Because morphogens are critical for initiating many patterning processes, there has been much interest in how they act at various distances to activate or suppress gene expression. We know that, in some cases, morphogens act over short distances (one to five cells); in others they influence cell fates over a much longer range (10–30 cells). Now, Ashe and Levine¹ (page 427 of this issue) provide the first evidence that a morphogen known as Short gastrulation (Sog) exerts different effects at short versus long range.

Sog is involved in patterning the dorsal–ventral axis of the early fruitfly embryo^{2–4} at what is known as the blastoderm stage, when all cells seem to be morphologically equivalent. It is thought to do so by opposing the action of another morphogen called Decapentaplegic (Dpp)^{3–5}, a secreted protein⁶ that belongs to the family of bone-morphogenetic proteins (BMPs). Sog is expressed in the lateral cells that comprise the neuroectoderm (which contributes to formation of the nervous system), whereas Dpp is expressed in adjacent dorsal cells. These cells give rise to the dorsal epidermis and to an extra-embryonic tissue called the amnioserosa, which forms at the dorsal extremity of the embryo (Fig. 1a). There is much evidence to suggest that peak levels of Dpp signalling specify the presumptive amnioserosa, whereas lower levels define the dorsal ectoderm⁷. Within the lateral neuroectoderm, Sog blocks Dpp signalling, allowing the cells there to follow their default preference of neural development^{5,8} (Fig. 1a). This action of Sog in early-blastoderm-stage embryos limits the spread of Dpp signalling into the neuroectoderm⁵, by a mechanism that has been highly conserved during evolution^{8,9}.

How does Sog suppress Dpp signalling asymmetrically, in the dorsal region of the embryo (Fig. 1b)? One model is that, in mid-blastoderm-stage embryos, Sog diffuses dorsally from the neuroectoderm and is degraded by Tollid (Tld) — a metalloprotease that is expressed only in dorsal cells¹⁰. The theory

goes that, by having a separated source and sink for Sog, a stable concentration gradient is created. Accordingly, the concentration of Sog is highest in cells immediately adjacent to the neuroectoderm (that is, the dorsal ectoderm cells) and lower in cells further from the source of Sog (the amnioserosa cells). This hypothetical Sog concentration gradient is imagined, in turn, to generate a reciprocal gradient of Dpp activity, which is maximal dorsally (in the amnioserosa) and lower ventrally (in the dorsal ectoderm). Consistent with the idea that Sog diffuses dorsally and antagonizes Dpp signalling, if expression of the *sog* gene is reduced, weak loss-of-function mutants in the Dpp pathway can be rescued^{3,4}. Reduced levels of *sog* also result in vertical displacement of the border between the dorsal ectoderm and the amnioserosa⁵, consistent with Sog being a long-range morphogen.

But this model does not explain why *sog*[−] mutants largely lack amnioserosa. Early *sog*[−] embryos do, in fact, have an expanded primordium for the amnioserosa, but these amnioserosa cells die prematurely². Perhaps cells in the domains that correspond to the future amnioserosa and dorsal ectoderm cells need to communicate, to refine and maintain the distinctions between cells created by different levels of Dpp signalling.

Alternatively, Sog might promote the formation of amnioserosa cells at a distance, as well as antagonizing Dpp signalling locally (Fig. 1c). The results of Ashe and Levine¹ argue in favour of this second possibility.

Ashe and Levine uncovered a potential long-range function for Sog by expressing the *sog* gene in a stripe along the anterior–posterior axis. This clever strategy uncouples the action of Sog from other, possibly redundant, dorsal–ventral patterning elements. When the authors examined expression of a marker for peak Dpp expression, they found that Sog had two distinct effects — it suppressed expression of the Dpp marker near the stripe of Sog expression, but broadened expression at a distance from the Sog source. So, Sog seems to exert opposing effects at short versus long range.

Ashe and Levine next used the same strategy to express Sog in mutant embryos with moderate Dpp signalling in all cells. They found that the Tld protease is critical for long-range activation by Sog, yet is not required for the short-range inhibitory effect. Consistent with the need for Tld in long-range activation, a membrane-tethered form of Sog was effective only at short-range inhibition of Dpp.

The new results raise several questions. First, is the long-range positive action of Sog direct, or is it propagated by a relay mechanism? Because the authors could uncouple long-range activation from short-range inhibition, the long-range activity is probably direct. Second, what is the nature of this activity? Perhaps Sog, or a fragment generated by Tld cleavage, forms a complex with Dpp, stimulating its activity. Or maybe a form of Sog acts directly on Dpp receptors to promote signalling. Alternatively, Dpp signalling would be stimulated if Sog inhibited a BMP-related molecule that normally

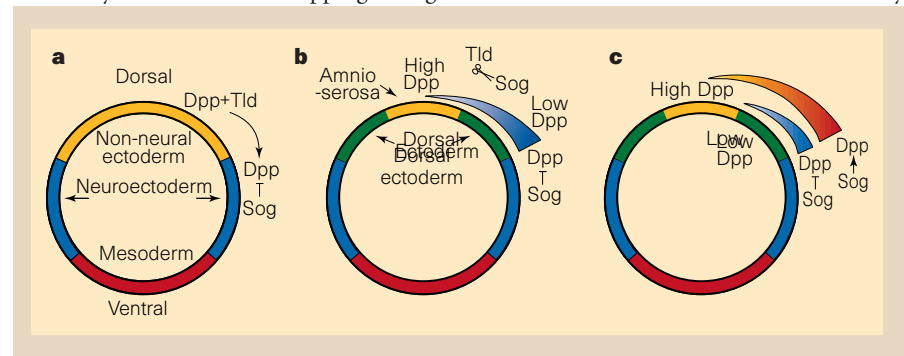


Figure 1 Signalling by Sog. a, Cross-sectional view of an early-blastoderm-stage embryo. Sog is expressed in lateral cells comprising the neuroectoderm; Dpp and Tld are expressed in dorsal cells comprising the non-neural ectoderm. Sog prevents Dpp signalling from spreading into the neuroectoderm by interfering with a positive-feedback loop created by Dpp diffusing and activating its own expression⁸. b, A mid-blastoderm-stage embryo, in which Sog signalling subdivides the dorsal region into a zone of high Dpp signalling (amnioserosa) and a zone of low Dpp signalling (dorsal ectoderm). The hypothetical Sog concentration gradient is created owing to the source of Sog in the neuroectoderm (high concentration) and its degradation by Tld in dorsal cells (low concentration). In this model, Sog only inhibits Dpp signalling. c, A mid-blastoderm-stage embryo in which Sog inhibits Dpp signalling locally and promotes Dpp signalling at long range, based on the results of Ashe and Levine¹.

reduces its activity. Third, why does Sog activate Dpp signalling only at a distance? Can alternative forms or complexes of Sog diffuse to a greater or lesser extent? Finally, why is the border between the domains with high Dpp activity, and those with low activity, so sharp? Does some kind of 'secondary' refinement contribute to the stable sub-division of the dorsal domain? It is clear that the Soga will continue. □

Ethan Bier is in the Department of Biology, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093-0349, USA.

e-mail: bier@biomail.ucsd.edu

1. Ashe, H. L. & Levine, M. *Nature* **398**, 427–431 (1999).
2. Zusman, S. B., Sweeton, D. & Wieschaus, E. F. *Dev. Biol.* **129**, 417–427 (1988).
3. Ferguson, E. L. & Anderson, K. V. *Development* **114**, 583–597 (1992).
4. François, V., Solloway, M., O'Neill, J. W., Emery, J. & Bier, E. *Genes Dev.* **8**, 2602–2616 (1994).
5. Biehs, B., François, V. & Bier, E. *Genes Dev.* **10**, 2922–2934 (1996).
6. Padgett, R. W., St Johnson, R. D. & Gelbart, W. M. *Nature* **325**, 81–84 (1987).
7. Rusch, J. & Levine, M. *Curr. Opin. Genet. Dev.* **6**, 416–423 (1996).
8. Bier, E. *Cell* **89**, 681–684 (1997).
9. Hemmati-Brivanlou, A. & Melton, D. A. *Cell* **88**, 13–17 (1997).
10. Marques, G. et al. *Cell* **91**, 417–426 (1997).

Microbiology

Millennium bug

R. S. Siew

*So, naturalists observe, a flea
Hath smaller fleas that on him prey;
And these have smaller fleas to bite 'em,
And so proceed ad infinitum.
Thus every poet, in his kind,
Is bit by him that comes behind.*

Jonathan Swift's ditty is true even for those smallest of parasites, the viruses. For example, adeno-associated virus is a parvovirus (parvo means small), which cannot replicate without its larger companion. Similarly, the hepatitis delta agent (an RNA genome related to plant viroids) depends on hepatitis B (a DNA virus) for encapsidation¹, the process in which the viral nucleic acid is packaged in protein.

In the era before antibiotics, phages were thought to be ideal viruses to control bacterial infections. Now evidence is emerging that a satellite RNA phage² with retrovirus-like properties first colonizes *Salmonella enteritidis* in chickens and then transmits its genome to cells of the chicken itself³. This agent — named bellophage after Hilaire Belloc, who came behind Swift to muse on microbes — could thus act as a shuttle vector moving gene sequences between bacteria and their animal hosts⁴. The findings have stimulated debate among microbiologists and evolutionary biologists, and are attracting the attention of both the biotechnology industry and environmentalists.

Bellophage has a tiny 1-kilobase genome encoding a nucleocapsid protein, a replicase component and an integrase². The replicase is minute but serves as a reverse transcriptase, and it was a puzzle how such a small protein could perform the complex polymerase and RNaseH functions. Recently it became apparent that, rather like the replicase of the RNA phage Q β , it binds to host DNA polymerase to change the enzyme's specificity so that the complex acts as an RNA-directed DNA polymerase⁵. The newly formed DNA genome then becomes inserted as prophage into the host genome with the aid of inte-

grase. The nucleocapsid also has intriguing properties. It is essentially a leucine zipper, a structural motif that features in many protein–protein interactions, with a carboxy-terminal basic domain which brings the bellophage RNA to associate with the capsid protein of its DNA helper phage, omega².

The claimants to the discovery of bellophage^{6,7} agree on its remarkable ability to adopt new helper viruses and indeed new hosts. Although the main reservoir appears to be omega-phage-carrying *Salmonella* in the gut of chickens, there is growing evidence that it can be packaged by adenovirus and even influenza capsid proteins (in birds, both of these viruses also persist as gut infections). Transcapsidation by the animal viruses, however, requires that the bellophage genome transfers from *Salmonella* to the gut epithelium itself. The capacity of the bellovirus, as we must now call it, to mobilize different helper viruses is correlated with hypermutation in the leucine zipper region of its nucleocapsid, generated by the lack of editing in the reverse transcription step in its life cycle. Moreover, a crucial Y2K mutation allows the provirus to remain latent in the host until its activation triggers programmed cell death⁸.

At a symposium in Sri Lanka, held last month, virologists and evolutionary biologists discussed whether bellovirus has newly emerged as an animal parasite. Free-living bacteria often exist in complex biofilms⁹ that also contain single-cell eukaryotes, such as algae and protozoa, and this may long have provided an opportunity for viruses to shuttle between bacterial, plant and animal host kingdoms. Mathematical biologists have become fascinated by bellophage¹⁰ — computer modelling demonstrates how the evolutionary dynamics of bellophage hypermutation yield strikingly different outcomes in *Salmonella* and chickens, the virus seemingly oscillating between extreme virulence and aggressive symbiosis.

As to which came first, so to speak, the

chicken or the phage, the answer has come from another quite unexpected feature of the bellovirus life cycle. In chicken cells the provirus does not integrate into nuclear DNA, but rather into mitochondrial DNA¹¹. The mitochondrial genome is of course related to bacterial DNA, especially of intracellular parasites such as *Rickettsia*¹². This is the first example of viral genome insertion into mtDNA as a kind of retrotransposon and it may explain why bellovirus can cross host kingdoms. Comparative studies of mtDNA reveal traces of bellovirus-related sequences in some host species but not in others. For example, integrase sequences have been detected in both human and chimpanzee mtDNA, but not in that of orang-utans¹³. This analysis of mitochondrial 'Eve' not only confirms that the Garden of Eden really was in Africa, but also indicates that bellovirus-like transposition has occurred among apes as well as chickens.

While academics ponder co-evolution versus genetic mobility, the biotechnology industry has been quick to see more profitable opportunities. A start-up company in Britain, Oxgal, is testing Bellovir, a nucleoside analogue which blocks bellophage reverse transcription¹⁴ (though preliminary data indicate that Bellovir also inhibits telomerase, resulting in feather loss in chickens treated with high doses of the drug). Another company, Tomannos, has patented a bellovirus variant that integrates into chloroplast DNA as a possible vector for genetically modified plants.

It is hardly surprising that public health officials and environmentalists will wish to keep a close watch on bellophage. Manipulation of such a versatile virus could allow it to escape in undesirable ways that with hindsight might make us appear foolish. Belloc may have foreseen this 101 years ago when, inspired by Samuel Butler¹⁵, he ended his poem, *The Microbe*:

*But scientists, who ought to know,
Assure us that this must be so.
Oh! let us never, never doubt
What nobody is sure about!* □

R. S. Siew is at the Windeyer Institute, University College London, 46 Cleveland Street, London W1P 6DP, UK.

e-mail: windinst@ucl.ac.uk

1. Carman, W. F. & Trautwein, C. in *HIV and the New Viruses* (eds Dalglish, A. G. & Weiss, R.) 415–460 (Academic, London, 1999).
2. Ffowl, A. *Phil. Trans. Acad. Sci.* **237**, 1951–1970 (1999).
3. Lagado, R. S. J. *Chem. Biol.* **88**, 10001–10006 (1998).
4. Chesterton, G. K. et al. *Protista* **6**, 123–130 (1997).
5. Nimet, M. H. & Inatuzim, S. J. *Enzymol.* **391**, 121–125 (1998).
6. Issounis, F. B. et al. *Avian Virol.* **41**, 525–535 (1997).
7. Ollag, C. R. et al. *Nature* **394**, 967–974 (1998).
8. Sisir, C. & Linnell, K. *Silico* (in the press).
9. Brown, M. R. W. & Barker, J. *Trends Microbiol.* **7**, 46–50 (1999).
10. Yam, R. & Kawon, M. *Eneics* **284**, 60–64 (1999).
11. Silugram, L. et al. *Plastid* **16**, 232–230 (1999).
12. Gray, M. W., Burger, G. & Lang, B. F. *Science* **283**, 1476–1481 (1999).
13. Prash, P. & Nhai, H. *Nature* **394**, 1999–2001 (1998).
14. Doh, D. & Eroon, J. *Tecnal* **335**, 871–872 (1998).
15. Butler, S. *Erewhon* (Trübner, London, 1872).

Synthetic chemistry

Sugars slide into heparin activity

Pierre Sinaj

The manipulation of carbohydrate chemistry provides, in principle, a way of building up polysaccharides of any complexity. An even greater challenge is how to use organic synthesis to explore molecular glycobiology and to devise, then produce, chemical architectures with potential medicinal and pharmaceutical importance. Probably the best illustration to date of such a powerful application of synthetic organic chemistry is described on page 417 of this issue by Petitou *et al.*¹ They have, for the first time, designed and chemically synthesized a complex oligosaccharide that is ten times more potent *in vivo* than both standard heparin and low-molecular-weight heparin, which are major drugs for the prevention and treatment of cardiovascular diseases. It is also remarkably devoid of certain detrimental side effects when compared with natural heparin.

The coagulation of blood is the result of a cascade of events employing various enzymes called activated blood-coagulation factors, one of the most important being thrombin. The protein antithrombin (AT) is the main physiological inhibitor of these factors and its action is amplified after binding to heparin. This is the origin of the anticoagulant and antithrombotic activities of heparin — an extraordinarily complex, natural polysaccharide, which typically carries negatively charged groups such as carboxylates and sulphates. Most of the numerous biological activities known to be associated with heparin are due to binding interactions between these anionic groups and the positively charged groups of a large variety of proteins.

A long-standing enigma in heparinology was whether the interaction between heparin and antithrombin was saccharide specific. A breakthrough came in the early 1980s with the recognition^{2,3} of a structurally unique sequence of five sugars within the polysaccharide, capable of highly selective and tight binding to the protein antithrombin (Fig. 1a). Organic synthesis then made its first contribution to the modern history of heparin: the total synthesis of this pentasaccharide was achieved⁴, allowing unambiguous confirmation of the specific nature of the interaction. Indeed, a synthetic tetrasaccharide (four-sugar sequence) does not bind to AT; in contrast, the synthetic pentasaccharide strongly binds to AT, forming an equimolar complex. The first generation of biologically active synthetic oligosaccharides was born.

In natural heparin, this specific pentasaccharide sequence (A-domain) is embed-

ded into a long, regular polymer (T-domain), made from a repeating disaccharide sequence armed with four negative charges (Fig. 1a). This part of the complex structure electrostatically attracts thrombin and then acts as a molecular slide, guiding it to a position where it collides with antithrombin, which is tightly bound onto

the A-domain of the polysaccharide. Thrombin is then inhibited, and therefore so is the blood coagulation cascade. It has also been suggested, through molecular modelling⁵ and crystallography studies⁶, that the sliding of thrombin on heparin operates from left to right.

From this point on, the ultimate goal would be to obtain a fully synthetic drug. The synthesis of the natural carbohydrate machinery, that is, the pentasaccharide A-domain, equipped with a long enough anionic slide (T-domain) on the left side, is a fascinating but formidable challenge. As a

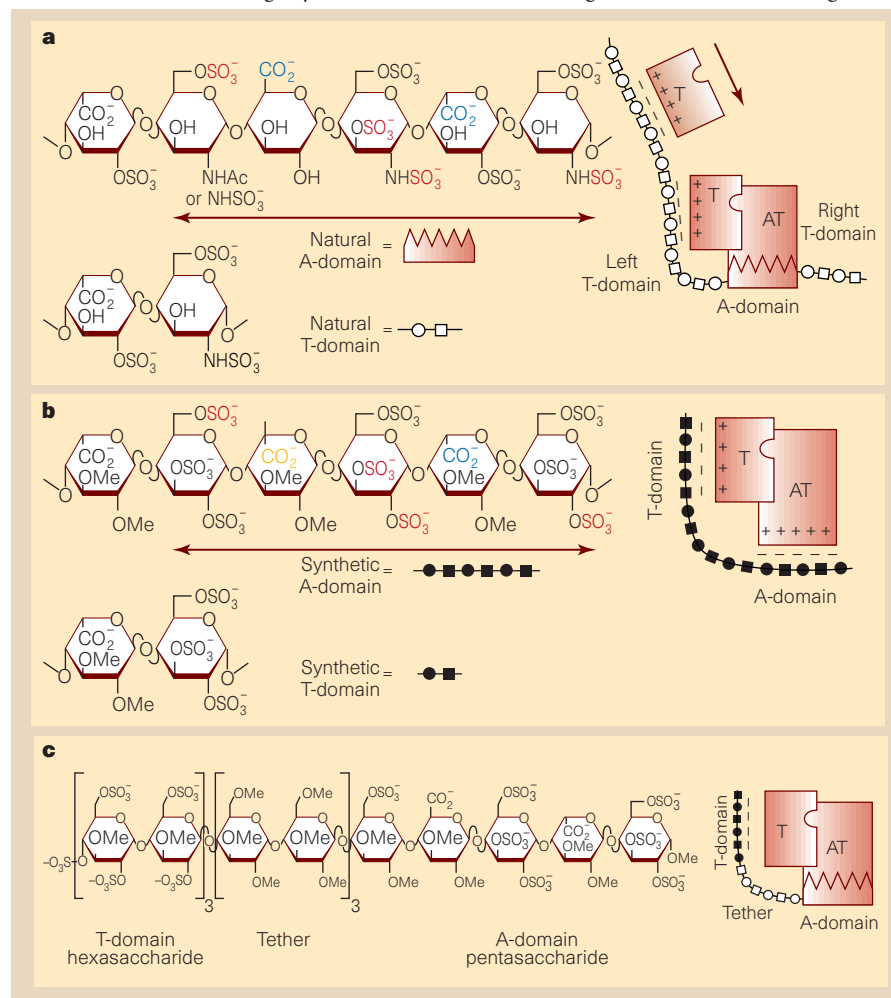


Figure 1 Natural and synthetic systems for inhibiting the enzyme thrombin (T). a, The physiological inhibitor antithrombin (AT) tightly binds onto a unique pentasaccharide sequence (A-domain) present in heparin. Six positively charged groups in this domain have been shown to be essential for selective and tight binding to AT: four sulphonate groups (red) and two carboxylate groups (blue).

Thrombin is electrostatically attracted by the regular part of heparin (T-domain); it slides along it until it hooks onto AT to inhibit thrombin and the blood-coagulation cascade. b, A regular hexadecasaccharide constructed by chemical assembly of a single disaccharide structure. This oligosaccharide machinery⁷ also results in the inhibition of thrombin. AT binds to any pentasaccharide sequence of this system, because such a domain contains the four essential sulphonate groups (red) and one carboxylate group (blue). The affinity is decreased compared with the natural pentasaccharide sequence, owing to the change in spatial orientation of the other carboxylate group (yellow). But this structural change allows AT to slide more freely along this system until it becomes bound in a position that mimics the situation occurring in natural heparin. c, The remarkable synthetic heptadecasaccharide machinery devised by Petitou *et al.*¹ is composed of three parts: a second-generation pentasaccharide mimicking the A-domain of natural heparin and having high affinity for AT; a regular anionic hexasaccharide as a T-domain; and a tether between these two domains made from a neutral methylated hexasaccharide.

first move, Petitou *et al.*⁷ made a critical step in their analysis of the process, then devised and chemically synthesized a more simplified system constructed from just a single disaccharide unit.

Two structural features along the A-domain of natural heparin are essential for tight binding to AT: the presence of four charged sulphate groups (Fig. 1a), and of two different acidic sugars (uronic acids). Taking this into account, a second generation of mimetics has been synthesized⁸, where O-sulphonato groups replace N-sulphonato substituents, and hydroxyl groups are methylated. These modifications fully preserve the specific and tight binding to AT, yet significantly simplify the chemical synthesis. However, such prototypes still appear relatively complex as synthetic targets, because the two uronic acids are different, which is an essential structural feature of high-affinity compounds.

Petitou *et al.*⁷ added together three identical disaccharide units to give a regular hexasaccharide containing only one type of uronic acid (L-iduronic acid), but otherwise having all the essential sulphate groups in the right position (Fig. 1b). This audacious simplification gave a compound that still binds to AT, although the affinity dropped two to three orders of magnitude lower than the natural pentasaccharide, or the first-generation synthetic mimics, and was in the micromolar range. This is exactly what Petitou *et al.* were looking for. This heavily sulphated structure is also able to bind thrombin with about the same affinity. In other words, appropriate oligomerization of a single synthetic disaccharide unit will produce a charged carbohydrate chain that acts as a slide for both thrombin and antithrombin. So it is no longer necessary to worry about the position of the A-domain in this machinery (left side or right side): AT is now free to move along the chain, resulting in a self-organizing complex, composed of AT, thrombin and this new third-generation prototype. It was found that synthetic slides containing up to 14 sugar units were inactive in the thrombin inhibition assay, whereas the activity of a hexadecasaccharide (containing 16 sugars) was demonstrated, and this antithrombin activity grew with increasing chain length. This work also elegantly solves a long-standing question⁹ about the minimum size of heparin fragments that can inhibit thrombin. It is truly remarkable that a key feature of this discovery comes from a deliberate and significant lowering of the affinity of two molecules, a move that is not intellectually instinctive in molecular glycobiology.

As elegant as it is, this synthetic endeavour produces drug candidates which do indeed have antithrombotic activity similar to natural heparin, but unfortunately suffer

the same side effects *in vivo* as a result of platelet activation. Ideally, active drugs should be able to discriminate between thrombin and other proteins, particularly platelet factor-4 (PF4), which is related to heparin-induced thrombocytopenia. In the search for such specificity, Petitou *et al.*¹ synthesized oligosaccharides containing the high-affinity A-domain (second-generation pentasaccharide) attached to a regular left-hand slide simply composed of disulphated glucose units. In accordance with previous results, the antithrombin activity of these oligosaccharides was size dependent and was present at (and above) 15 saccharide residues. A compound with 19 saccharide residues was as potent as the most active fraction isolated from a natural heparin preparation. But again, reducing the size of the natural heparin fragment down to the minimum that allowed thrombin inhibition, was not sufficient to abolish the undesirable anionic interaction with PF4.

Petitou and co-workers have now made the ingenious final move¹. They take into account that the minimum size of the anionic thrombin slide (T-domain) required for activity is a tetra- or hexasaccharide¹⁰ and that an octasaccharide, or larger, fragment is required for PF4 binding¹¹; also, as previously demonstrated by organic synthesis, the minimum length of a significantly active antithrombin oligosaccharide is 17 residues. The result is the novel synthetic machinery shown in Fig. 1c, which is a heptadecasaccharide (17 sugar units), containing a simplified pentasaccharide A-domain with high affinity for antithrombin at the right end, and a regular anionic hexasaccharide as a T-domain at the left end. The connecting hexasaccharide used to mimic the natural carbohydrate skeleton of heparin need not be sulphated, and therefore a readily available neutral oligosaccharide was used. Hopefully, PF4 will not be able to bind onto this molecule. Indeed, such a synthetic

compound is not only unable to activate platelets in the presence of plasma from patients with heparin-induced thrombocytopenia, but is also devoid of other side effects attributed to the anionic charges of heparin, such as haemorrhaging.

This enlightening study demonstrates the power of chemical organic synthesis in the field of heparinology and clearly offers new perspectives on heparinotherapy. For the first time, undesirable effects of natural heparin have been eradicated through structural changes. Whatever the final outcome of clinical trials, this work also greatly enhances our present understanding of the structural role of heparin in biology.

Heparin controls a large variety of biological processes by interacting with proteins, such as growth factors. So far, apart from the unique pentasaccharide sequence that tightly binds AT, no other specific sequence having selective and tight affinity for a protein has been identified. Therefore, it is not unreasonable to speculate that, in the future, multistep-convergent synthesis will play a central part in understanding the nature of the numerous structural features of heparin. □

Pierre Sinay is in the *Departement de Chimie, Ecole Normale Supérieure, URA 1686, 24 rue Lhomond, 75231 Paris Cedex 05, France.*

e-mail: Pierre.Sinay@junie.ens.fr

1. Petitou, M. *et al.* *Nature* **398**, 417–422 (1999).
2. Choay, J. *et al.* *Ann. NY Acad. Sci.* **370**, 644–649 (1981).
3. Thunberg, L., Bäckström, G. & Lindahl, U. *Carbohydr. Res.* **100**, 393–410 (1982).
4. Choay, J. *et al.* *Biochem. Biophys. Res. Commun.* **116**, 492–499 (1983).
5. Grootenhuys, P. D. J., Westerduin, P., Meuleman, D., Petitou, M. & van Boeckel, C. A. A. *Nature Struct. Biol.* **2**, 736–739 (1995).
6. Jin, J. *et al.* *Proc. Natl Acad. Sci. USA* **94**, 14683–14688 (1997).
7. Petitou, M. *et al.* *Angew. Chem. Int. Edn* **37**, 3009–3014 (1998).
8. Basten, J. *et al.* *Bioorg. Med. Chem. Lett.* **2**, 905–910 (1992).
9. Laurent, T. C., Tengblad, A., Thunberg, L., Höök, M. & Lindahl, U. *Biochem. J.* **175**, 691–701 (1978).
10. Olson, S. T., Halvorsen, H. R. & Björk, I. *J. Biol. Chem.* **266**, 6342–6352 (1991).
11. Maccarana, M. & Lindahl, U. *Glycobiology* **3**, 271–277 (1993).

Sonoluminescence

And there was light!

Robert Apfel

It took ten years, but on page 402 of this issue Hilgenfeldt *et al.*¹ are able to explain the tiny dot of light emanating from a solitary, sonically driven bubble, first observed by Felipe Gaitan in 1989. Gaitan had been adjusting the parameters in his apparatus to study a new regime for acoustically levitating a bubble in water — a technique perfected almost three decades before by his PhD advisor, Lawrence Crum — and was surprised to find the bubble emitting what appeared to be a continuous dot of light. Light from sound, or sonoluminescence, was not a new

phenomenon, but never had a single bubble been controlled in such a manner. In fact, after careful observation with an appropriate photodetector Gaitan determined that the light was not in fact continuous, but emitted in an extremely brief burst each acoustic cycle^{2,3}. That's 20,000 flashes per second, or 72 million synchronous flashes of light every hour.

During a visit to Crum's laboratory, Seth Putterman immediately realized that this remarkable phenomenon, called single-bubble sonoluminescence (SBSL), might

represent new and exciting physics. He set about quantifying with great precision the remarkable characteristics of this light emission, finding that the lifetime and the pulse-to-pulse synchronicity of each flash was on the order of 50 trillionths of a second. His group went on to thoroughly examine how SBSL was dependent on a wide range of experimental parameters⁴. Over 100 peer-reviewed papers have been published on SBSL since its discovery by Gaitan, but the mechanism of light emission remained elusive until now.

Is this new physics, as Putterman originally thought? The answer from Hilgenfeldt *et al.*¹ is no. But it took some pretty fancy detective work to resolve paradoxes and integrate physical principles — over enormously disparate ranges of energy and time — to come up with an illuminating explanation that quantitatively matched the data.

The broad understanding of the phenomenon is now as follows: the rarefied phase of the sound wave grows a bubble from about 5 μm in diameter to about 70 μm in diameter (around the thickness of a human hair). Then, as the sound field turns compressive, the virtually empty spherical bubble comes crashing in on itself, powered by the inertia of the surrounding water. The bubble maintains its spherical symmetry for almost the entire collapse phase, therefore optimizing the internal energy concentration. As the collapse accelerates, the residual gas caught inside the bubble is rapidly compressed, raising its temperature to 20,000–30,000 K. These temperatures are sufficient to create a plasma of ions, neutral atoms and electrons. By keeping track of the ways in which thermal energy in the collapsing bubble is distributed among these constituents, one can then understand the reverse process of light emission and extinction as the energy density of the expanding bubble is reduced and photons are emitted. About 20 μs later the processes of growth, collapse and light emission repeat themselves, resulting in a pulse of light every acoustic cycle.

The issue of new physics raised by Putterman is almost overshadowed by the fact that this ever-flashing bubble appears to be a wonderful laboratory for acoustics, fluid dynamics, thermodynamics, optics, plasma and quantum physics, all under highly dynamic conditions. And because of the high temperatures and pressures achieved inside the bubble, there is also some remarkable chemistry going on when diatomic gases are present. In what is now called the 'dissociation hypothesis'⁵, diatomic gases (such as nitrogen) actually dissociate and dissolve into the surrounding fluid during the bubble collapse, leaving only noble gases (such as argon). Further experiments^{6–8} have provided conclusive evidence for the validity of this hypothesis.

Hilgenfeldt *et al.* have now been able to tie this chemistry and physics together to quantitatively explain many of the observations made for SBSL.

What is both remarkable and limiting in the sonoluminescence phenomenon is that the stability required to produce the clock-like synchronicity of SBSL depends largely on the behaviour of the bubble while it is relatively large, whereas the light emission occurs when the bubble is close to its smallest size. This coupling of bubble stability and light emission has discouraged researchers from going 'outside the box' to optimize the bubble growth, so as to maximize the internal temperature upon collapse.

Why increase the internal bubble temperature? If one were to grow the bubble by a factor of ten or more — to 1 mm in diameter — what might the temperature inside the bubble reach at its collapse? Could it reach the conditions that would allow for nuclear (hot) fusion inside the bubble if deuterium gas or a combination of deuterium and tritium gases were inside the bubble? No one can say for sure whether this is possible, and a small community of sonoluminescence researchers is quietly going on with this work, trying to avoid the hype that accompanied the 'cold' fusion claims of ten years ago. If it were possible, this would be table-top micro-thermonuclear fusion as suggested by Willy Moss of Lawrence Livermore Laboratory, whose computations helped to clarify the optical emission mechanism of SBSL⁹. Moss continues to provide guidance to experimentalists searching for what he calls a 'star in a jar'.

And there is another reason for the sustained interest in SBSL. Put simply, this bubble burning bright is 'cool' and accessible science. Here is a most amazing reactor for studying basic physics. Imagine if one could actually predict the onset conditions for the emission of neutrons from a collapsing bubble, in a jar of degassed heavy water through which an audible sound wave was propagating — it would be a resounding affirmation of the unity of physics, and we may not have to wait another ten years. □

Robert Apfel is in the Department of Mechanical Engineering, Yale University, 9 Hillhouse Avenue, New Haven, Connecticut 06520-8286, USA.
e-mail: robert.apfel@yale.edu

1. Hilgenfeldt, S., Grossmann, S. & Lohse, D. *Nature* **398**, 402–405 (1999).
2. Gaitan, D. F. thesis, Univ. Mississippi (1990).
3. Gaitan, D. F., Crum, L. A., Church, C. C. & Roy, R. A. *J. Acoust. Soc. Am.* **91**, 3166–3183 (1992).
4. Barber, B. P., Hiller, R. A., Löfstedt, R., Putterman, S. J. & Weninger, K. R. *Phys. Lett.* **281**, 66–143 (1997).
5. Lohse, D. & Hilgenfeldt, S. *J. Chem. Phys.* **107**, 6986–6997 (1997).
6. Holt, R. G. & Gaitan, D. F. *Phys. Rev. Lett.* **77**, 3791–3794 (1996).
7. Matula, T. J. & Crum, L. A. *Phys. Rev. Lett.* **80**, 865–868 (1998).
8. Ketterling, J. A. & Apfel, R. E. *Phys. Rev. Lett.* **81**, 4991–4994 (1998).
9. Moss, W. C. *et al. Phys. Rev. E* **59**, 2986–2992 (1999).

Daedalus

Oscillating enzymes

Last week Daedalus proposed piezoelectric catalysts, whose surfaces or cavities could be set into intense controlled vibration. Reagent molecules were adsorbed during the expansion phase; contraction brought them together to induce reaction; on the subsequent re-expansion the product molecule was set free.

He now muses that nature may have got there first. Enzymes, those amazingly specific biological catalysts, have extremely subtle surfaces. Only one molecule, or pair of molecules, can fit on a given surface; and their resulting reaction is extremely well defined. Daedalus reckons that the vibrational modes of the enzyme molecule must play a crucial part in its action.

All molecules, of course, can be excited into vibration by radiation of the right frequency. But only quite low-frequency modes, corresponding to wavelengths longer than about 0.05 mm in the far infrared, can be significantly excited merely by ambient temperature. And this, says Daedalus, explains why so many enzymes are such needlessly complicated molecules, very large compared with their active catalytic sites. An enzyme molecule has to be big and heavy, so as to have vibrational modes low enough to be usefully excited at body heat.

This train of thought has sparked two new DREADCO projects. The first and simpler is to irradiate yeast fermentations, penicillin production, sugar inversion and other well-known enzyme reactions, with intense beams of tuned far-infrared and millimetre-wave radiation, looking for sudden rises in output or changes of product. The active frequency will reveal the key vibrational mode of the enzyme. If enzyme reactions can indeed be accelerated or directed by well-judged irradiation, biotechnology will be transformed.

The second project is the creation of 'stripped down' enzyme molecules, equipped with active catalytic sites but without the surrounding heavy protein chains. Their vibrational modes will be much higher, in the mid- and near-infrared. When irradiated at these higher frequencies, they will churn out the product a lot faster; and they can be switched off simply by turning off the radiation. Again, the biotechnology industry should snap up the technique. Daedalus is also devising special diathermy machines and infrared health lamps, cunningly tuned to boost key biochemical reactions in their patients.

David Jones

Obituary

Gertrude Belle Elion (1918-99)

Pioneer of drug discovery

Trudy Elion was one of only half a dozen women to receive a Nobel Prize for Physiology or Medicine this century. But although this pioneer of rational drug design was proud of her achievements, she did not care for pomp and circumstance. What she really cared about was good science, which she believed would yield effective treatments for disease. Between her retirement in 1983 and her sudden death on 21 February 1999, aged 81, she never stopped working — in fact, few people would have been aware that she had officially retired.

Elion was born in 1918 in New York, from Lithuanian and Russian parentage. Her grandfather and mother both died of cancer while she was a teenager — events that were to shape her outlook. Having earned a bachelor's degree from Hunter College in 1937, and a master's degree in chemistry from New York University in 1941, she tried to get work as a scientist. But women were rarely seen in laboratories at that time. After taking jobs teaching nurses, and testing pickles and berries for a food company, Elion finally had her opportunity when the Second World War created a shortage of labour in the pharmaceutical industry. In 1944 she got a job at Burroughs Wellcome as an assistant in the lab of George Hitchings, who was to become her fellow Nobel laureate. At the time of her retirement, she held the post of head of the Department of Experimental Therapy, a department she created. From 1983, until her death, she was scientist emerita at Glaxo Wellcome in Research Triangle Park, North Carolina.

Together with Hitchings, Elion began by comparing normal human cells with cancer cells, protozoa, bacteria and viruses. Their aim was to pinpoint differences in how the nucleic acids are metabolized in these various cells. By exploiting those differences, the pair developed highly targeted drugs that selectively blocked the growth and replication of certain cancer cells and pathogens. This new approach to drug development was to revolutionize the management of many diseases.

In the 1940s, little was known about the synthesis of nucleic acids, beyond the fact that purines and pyrimidines are incorporated into them. To identify the pathways by which nucleic acids are built — and work out how these routes could be selectively blocked — Elion and Hitchings studied the bacterium *Lactobacillus casei*. By 1948 they had synthesized a compound which, by interfering with nucleic-acid

metabolism, inhibited growth of *L. casei*. This substance, known as diaminopurine, worked by incorporating itself into newly synthesized DNA chains in place of adenine. Although development of this molecule was eventually abandoned because of its toxic side effects, over the following three years Elion and Hitchings identified two successors — thioguanine and 6-mercaptopurine. In 1953, when 6-mercaptopurine was put into clinical trials for the treatment of childhood leukaemia, it resulted in complete remission in one in three patients.

Elion and Hitchings' approach was revolutionary. They used their understanding of the structure of nucleic acids to synthesize molecules to specific targets (rather than simply screening randomly chosen molecules, which was more common at that time). This approach yielded new drugs at an extraordinary rate. In 1950 came pyrimethamine, a drug with a strong affinity for the enzyme dihydrofolate reductase. Pyrimethamine was shown to be 2,000 times more toxic to the malaria parasite than to its human host, and as such was used successfully in the management of malaria. In 1956 an antibacterial drug, trimethoprim, arrived, closely followed by azathioprine in 1957. Allopurinol, which blocks the formation of uric acid and so overcomes gout, was discovered six years later. And in 1977 Elion's group was prominent in the discovery of acyclovir, a drug that selectively blocks the replication of herpesvirus. The Nobel committee said in 1988, when it awarded the prize, that each of the drugs was worth a prize in itself.

The success of acyclovir overturned the assumptions of many biochemists — including, at one time, even Hitchings — who believed that it would be impossible to discover effective and selective antiviral agents. Elion's persistence was key to proving these assumptions wrong. She later wrote that this change of outlook was important in preparing the pharmaceutical industry to respond to the challenge of the human immunodeficiency virus (HIV). In fact, it was researchers trained by Elion who first saw the anti-HIV potential of 3'-azido-3'-deoxythymidine (AZT), at the time an unused anti-cancer drug from the 1960s.

Under Elion, the Department of Experimental Therapy at Burroughs Wellcome grew to more than 50 people. It was a diverse group, including chemists, biochemists, virologists and oncologists, and it would prove to be a model for the future. Although

Elion never finished her doctoral work, she was awarded 25 honorary degrees and many other honours. She also served on advisory committees for bodies as diverse as the World Health Organisation and the American Cancer Society.

Even after she had formally retired, Elion's energetic figure was regularly seen walking through the labs, listening to and talking with her colleagues. But she was always a few steps ahead of their thinking on new data, gently challenging their ideas and trying to ensure that there was good rationale for their science. She was often heard asking: "If we carry out these experiments, how will we use the information generated, and where will this lead us?"

Elion's innovative ways of working and thinking are as much a part of her legacy as the drugs that she discovered and the 45 patents on which her name appears. Indeed, she trained and mentored two generations of scientists; students were swept up in her passion to pursue science and medicine; and children were inspired by her infectious smile and boundless curiosity. This is Trudy Elion's real legacy, and one with which she would be very pleased.

Richard Kent and Brian Huber

Richard Kent and Brian Huber are in Research and Development, Glaxo Wellcome Inc., Five Moore Drive, Box 13398, Research Triangle Park, North Carolina 27709, USA.
e-mails: rk11377@glaxowellcome.com
beh13954@glaxowellcome.com



GLAXO WELLCOME

Angiogenesis inhibited by drinking tea

Consumption of tea has been shown to inhibit the growth of several tumour types in animals, including cancers of the lung and oesophagus¹⁻³. Drinking tea, especially green tea, is also associated with a lower incidence of human cancer⁴. The mechanisms of cancer inhibition are not known, although several hypotheses have been proposed. We investigated whether drinking green tea could suppress angiogenesis, a process of blood-vessel growth required for tumour growth and metastasis. We find that green tea, and one of its components, epigallocatechin-3-gallate (EGCG), significantly prevents the growth of new blood vessels in animals. This finding indicates that drinking tea may be beneficial for the prevention and treatment of angiogenesis-dependent diseases, including cancer and blindness caused by diabetes.

EGCG has been reported to inhibit urokinase, which may be used by some tumours for the invasion of neighbouring healthy tissues⁵. However, the effective concentration of EGCG seems to be too high for it to be physiologically relevant for tea

drinkers⁵. The growth of almost all types of tumour is dependent on angiogenesis, and the enlargement and metastasis of tumours is disrupted when this is suppressed⁶.

We thought that the inhibition of tumour growth by green tea might be mediated through the suppression of blood-vessel growth. To determine whether green tea could inhibit endothelial cell growth, we assayed EGCG on bovine capillary endothelial cells stimulated with the fibroblast growth factor FGF-2, as described⁷. We found that EGCG inhibited endothelial cell growth in a dose-dependent manner (Fig. 1a). The inhibition appeared to be specific to endothelial cells because non-endothelial cells, including murine T241 fibrosarcoma tumour cells, murine fibroblast cells, and rat smooth-muscle cells, were insensitive to EGCG treatment at the concentrations used.

We examined the effect of EGCG on angiogenesis in the chick chorioallantoic membrane assay⁸. Over a concentration range of 1 to 100 µg per disc, EGCG inhibited new blood-vessel growth in a dose-

dependent manner (Fig. 1g), as measured by the formation of avascular zones.

To investigate whether tea could suppress angiogenesis, we made green tea the sole drinking fluid for mice and examined the effect of oral consumption on the inhibition of corneal neovascularization stimulated by vascular endothelial growth factor (VEGF). The corneal model is a rigorous antiangiogenic assay, requiring systemic administration of a putative angiogenesis inhibitor to suppress neovascularization induced by 160 ng of VEGF in the cornea. The amount of green tea in the drinking water was 1.25% (4.69 mg ml⁻¹) containing 708 µg ml⁻¹ EGCG^{9,10}. The concentration of EGCG in the plasma was previously reported to be in the range of 0.1–0.3 µM, which is similar to levels in humans after drinking two or three cups of tea¹¹.

This tea preparation has been shown to significantly suppress the growth and progression of tumours in the lungs in animal models⁹. Compared with the control group that drank water alone (Fig. 1b), drinking tea significantly prevented VEGF-induced corneal neovascularization (Fig. 1c). Blood-vessel length (Fig. 1d), clock-hours of corneal neovascularization (the proportion of the circumference that is vascularized if the eye is viewed as a clock; Fig. 1e) and area of neovascularization (Fig. 1f) in seven corneas from four mice in the tea-drinking group were inhibited by approximately 55, 35 and 70%, respectively.

Our data indicate that EGCG suppresses endothelial cell growth *in vitro* and the formation of new blood vessels in chick chorioallantoic membrane. Drinking green tea significantly prevents corneal neovascularization induced by one of the most potent angiogenic factors, VEGF. Because the growth of all solid tumours is dependent on angiogenesis⁶, this finding may explain why drinking green tea prevents the growth of a variety of different types of tumour.

Yihai Cao, Renhai Cao

Laboratory of Angiogenesis Research,
Microbiology and Tumor Biology Center,
Karolinska Institute,
171 77 Stockholm, Sweden
e-mail: yihai.cao@mtc.ki.se

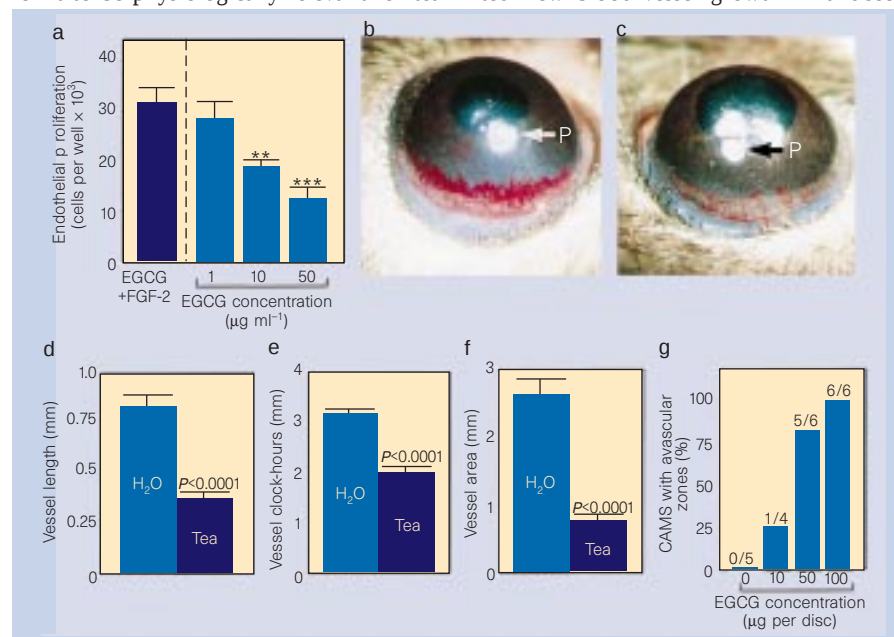


Figure 1 Suppression of endothelial growth and angiogenesis by EGCG and green tea. **a**, Results of a bovine capillary endothelial cell proliferation assay⁷. We seeded 10,000 cells per well in gelatinized 24-well plates and grew them in DME medium containing 5% bovine calf serum. EGCG samples in triplicates and FGF-2 at a final concentration of 1 ng ml⁻¹ were added to each well. After 72 h incubation, cells were counted with a Coulter counter. Values are mean (± s.e.m.) number of cells from 3 determinations. ***P* < 0.005, ****P* < 0.0001. **b–f**, Suppression of VEGF-stimulated vascularization by green tea in the mouse corneal model. Mice drank Chinese green-tea water extract^{9,10} starting at day 3 before VEGF implantation. Pellets containing 160 ng VEGF, sucrose, aluminium sulphate and hydron were implanted into corneal micropockets⁸. **b**, Cornea of a water-drinking mouse 5 days after implantation. **c**, Cornea of a green-tea-drinking mouse. **d–f**, Vessel length, clock-hours and vessel area. **g**, Results of the chorioallantoic membrane (CAM) assay⁸. Discs of methylcellulose containing EGCG dried on nylon mesh were implanted on the CAMs of individual embryos. Embryos and CAMs were examined by stereoscope after 48 h for the formation of avascular zones in the field of the implanted discs. At various doses, the number of CAMs with avascular zones over the total number of CAMs is indicated above each bar.

- Yang, C. S. & Wang, Z.-Y. *J. Natl Cancer Inst.* **85**, 1038–1049 (1993).
- Wang, Z.-Y. *et al. Carcinogenesis* **16**, 2143–2148 (1995).
- Yang, G.-Y. *et al. Cancer Res.* **57**, 1889–1894 (1997).
- Jankun, J., Selman, S. H. & Swiercz, R. *Nature* **387**, 561 (1997).
- Yang, C. S. *Nature* **389**, 134–135 (1997).
- Folkman, J. *Nature Med.* **1**, 27–31 (1995).
- Cao, Y. *et al. J. Biol. Chem.* **271**, 29461–29467 (1996).
- Cao, Y. *et al. J. Exp. Med.* **182**, 2069–2077 (1995).
- Yang, C. S. *et al. Exp. Lung Res.* **24**, 629–639 (1998).
- Wang, Z.-Y. *et al. Cancer Res.* **52**, 1162–1170 (1992).
- Yang, C. S. *et al. Cancer Epidemiol. Biomark. Prev.* **7**, 351–354 (1998).

Trimethylamine induces migration of waterfleas

It has long been known that the waterflea *Daphnia hyalina* exhibits diel vertical migration in the water column, but the chemical that triggers this behaviour has not been identified. We find that trimethylamine (TMA), which is a major component of the odour produced by decaying fish, induces *Daphnia* to migrate to greater depths during the day, presumably to avoid predation by fish^{1,2}. We observed a gradual increase in average depth of *Daphnia* with increasing TMA concentration. Changes in light intensity are known to trigger migration³, and chemicals produced by their predators must also be present⁴. Because migration has demographic and physiological costs, this chemical cue ensures that zooplankton migration occurs only when fish are present.

It has been shown that, whatever the migration trigger substance may be, it is produced only by fish in the presence of bacteria. This led to the hypothesis that trimethylamine-*N*-oxide (TMAO), which is important for cell volume regulation⁵ and protein stabilization⁶ in marine and freshwater fish, could be a precursor of the kairomone. Surplus TMAO is deposited in fish scales⁷, and bacteria of the skin mucus layer are known to reduce TMAO to TMA⁸.

We tested the hypothesis that TMA induces vertical migration in *Daphnia* by using a range of different concentrations, from 0 to 500 μM TMA, in autoclaved lake water⁹ (Fig. 1). Five individuals of a single *Daphnia hyalina* clone were put into each of four Perspex tubes (1 m long and 1.5 cm in

diameter) per treatment. The tubes were then placed in a water bath with a temperature gradient from top to bottom of 20 to 10 °C. We used a spectrophotometric assay procedure¹⁰ to establish the TMA concentration in 10 litres of the medium that had previously contained four fish (ides, *Leuciscus idus*, 10 cm long), and recorded the resulting vertical migration behaviour of daphnids exposed to this medium. TMA concentration in this 'fish' water was between 10 and 25 μM .

We found that even low concentrations of TMA induce *Daphnia* to migrate to deeper waters during the day in our test system. At night, they migrate back to the surface when exposed to small amounts of TMA, and stay near the bottom when TMA levels are high (more than 100 μM). When the lowest TMA concentration that induces vertical migration is compared with the activity in the fish water, it is clear that the reaction of *Daphnia* to fish water is stronger than just to TMA alone. This indicates that, although TMA is an active component of the 'fish factor', it is likely to be part of a cocktail of substances that deter *Daphnia*. The other substances probably help to reduce the chemical threshold that induces migration. It is not clear whether the TMA concentrations used in this study represent realistic values for aquatic systems, as we could find no published details of TMA concentrations in aquatic systems.

It has been shown⁹ that the fish kairomone is broken down by bacteria. We therefore compared the average day depth of daphnids exposed to 75 μM TMA in autoclaved water, and with the antibiotic ampicillin added and the TMA dissolved in non-autoclaved lake water. The migration activity of the *Daphnia* in non-sterile water slowly decreased over time (Fig. 2). After 72

hours, the average day depth of the animals in non-sterile conditions was no longer significantly different from the average day depth of animals in sterile control medium.

To investigate whether simply adding any substance to the water induces vertical migration, we added the same amount of TMAO and triethylamine to sterile water (both at 75 μM). As with the control *Daphnia*, there was no significant increase in day depth. We therefore conclude that the reaction to TMA is a specific one, and not simply a response to a change in conductivity or ionic strength of the medium.

Hinnerk Boriss, Maarten Boersma, Karen H. Wiltshire

Max-Planck-Institut für Limnologie,
Postfach 165, 24302 Plön, Germany
e-mail: boriss@mpil-ploen.mpg.de

1. Stich, H. B. & Lampert, W. *Nature* **293**, 396–398 (1981).
2. Gliwicz, M. Z. *Nature* **320**, 746–748 (1986).
3. Ringelberg, J. J. *Mar. Biol. Assoc. UK* **75**, 15–25 (1995).
4. Larsson, P. & Dodson, S. *Arch. Hydrobiol.* **129**, 129–155 (1993).
5. Lange, R. & Fugelli, K. *Comp. Biochem. Physiol.* **25**, 283–292 (1965).
6. Yancey, P. H. & Somero, G. N. *Biochem. J.* **183**, 317–323 (1979).
7. Hill, R. W. *Comparative Physiology of Animals* (Harper & Row, New York, 1976).
8. Gram, L., Wedell Neergaard, C. & Huss, H. H. *Int. J. Food Microbiol.* **10**, 303–316 (1990).
9. Loose, C. J., von Elert, E. & Dawidowicz, P. *Arch. Hydrobiol.* **126**, 329–337 (1993).
10. Kakác, B. & Vejtelek, Z. *J. Handbuch der Photometrischen Analyse Organischer Verbindungen* (Chemie, Weinheim, 1974).

Ancient stunted trees on cliffs

An undisturbed ancient woodland, dominated by tiny, slow-growing and widely spaced trees, grows on vertical cliffs of the Niagara escarpment in southern Canada¹. To investigate whether this woodland is unusual or is part of a previously undetected global pattern, we sampled ages and radial growth rates for trees on cliffs in the United States and in western Europe. We find that vertical cliffs often support populations of widely spaced trees that are exceptionally old, deformed and slow growing. Some of the most ancient and least-disturbed wooded habitat types on Earth are found on cliffs, even at sites close to heavy agricultural and industrial development.

We selected 21 cliffs for sampling in 15 eastern states of the United States, and 25 cliffs in eastern and southern Germany, eastern and southern France, central and northern England and Wales. All areas have been heavily developed for industry and agriculture over a long time. The cliffs visited were at least 500 metres long and between 20 and 1,500 metres high. We extracted increment cores for analysis from near the base of 224 mature trees. Some individual living trees grew as thin stem-strips on large amounts of dead wood. We

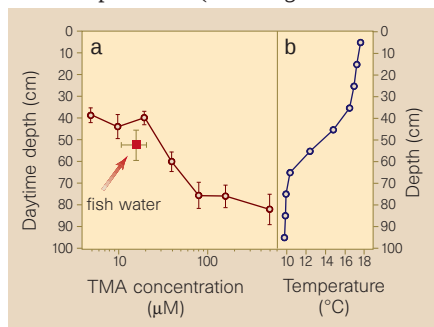


Figure 1 Behavioural responses of *Daphnia hyalina* to trimethylamine. **a**, Average day depth ($\pm$ standard error) of five daphnids (four replicates per treatment) on three consecutive days of different TMA concentrations. The day:night cycle was 16:8 hours. The reaction of the daphnids is plotted against the concentration of TMA in fish water. Bacterial decay of TMA was prevented by the addition of ampicillin. Migration depth increased in all tubes from day 1 to day 3. **b**, Temperature profile of the tubes. As TMA concentration increases, daphnids migrate to greater depths and colder water during the day.

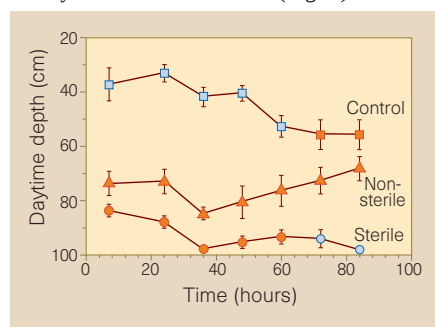


Figure 2 Average day depth ($\pm$ s.e.) of five daphnids. There were eight replicates per treatment. Symbols with different colours indicate significant differences between the treatments (Tukey post hoc comparison after analysis of variance, with treatment as a factor and time as a repeated measurement factor). Sterile and non-sterile treatments are not significantly different from each other, but were significantly different from the control for the first 60 hours; after 72 hours, daphnids kept under sterile conditions are significantly deeper than the control animals and those kept under non-sterile conditions.

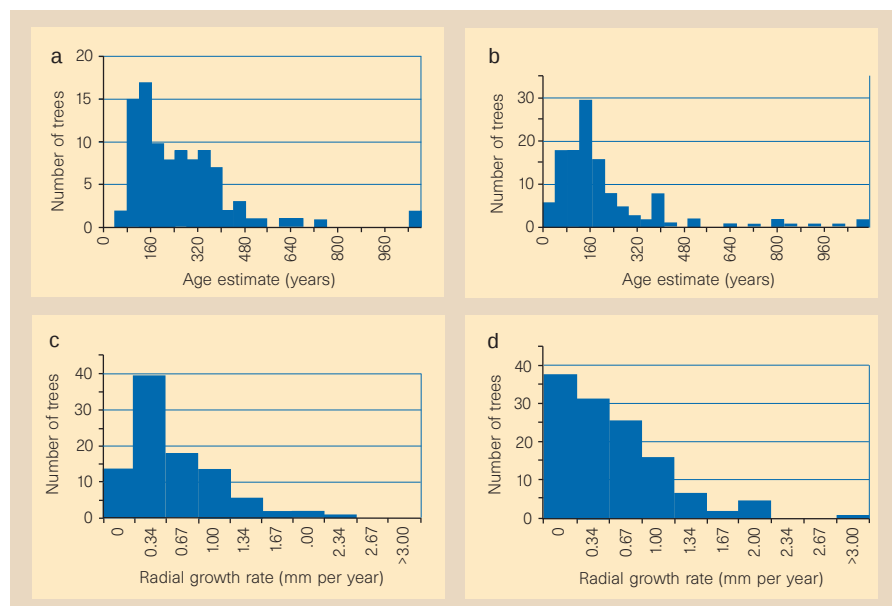


Figure 1 Age estimates and growth rates of trees on cliffs. Frequency distributions of estimated tree ages (a, b) and radial growth rates (c, d) for mature trees growing on cliffs in the United States (a, c, $n=97$) and Europe (b, d, $n=127$).

sometimes took cross-sections from the dead parts of these trees to allow for more accurate dating of core samples extracted from the living portions.

We recorded the diameter of each tree and the location of the pith for more accurate age estimations when core samples did not include the pith². We then used the distance from the cambium to the youngest ring to calculate radial growth rate, expressed as millimetres of new wood per year. All sample cores or sections were mounted and sanded³. Results were plotted as frequency distributions for both age estimate and growth rate, and US and European results were compared using non-parametric tests^{4,5}. For age estimates, we expected to find evidence of a negative exponential frequency distribution if the mature trees on the cliffs were part of a self-maintaining natural woodland⁶. For growth rate, we expected to find a tight clustering of very low values if the mature trees were highly constrained in their productivity⁷.

Most of the US trees sampled are between 160 and 400 years old, with two trees being more than 1,000 years old and very few less than 100 years old. The trees were mainly *Thuja occidentalis* in the northern states and *Juniperus virginiana* in the rest. Despite our non-random sampling, the portion of the curve above 80 years of age has the shape of a negative exponential (Fig. 1a). European trees also have a negative exponential age distribution when the youngest age category is excluded (Fig. 1b), and some *J. phoenicea* and *Taxus baccata* trees in France and the United Kingdom, respectively, were more than 1,000 years old. We found unevenly aged living tree populations on all cliffs, and observed

coarse woody debris in a variety of decay states. The findings indicate that the populations on both continents have remained undisturbed for long periods, and have had continuous, rather than pulsed, recruitment over the past millennium.

Radial growth rates for trees in the United States (Fig. 1c) and Europe (Fig. 1d) were mostly less than 1 millimetre per year, similar to the rate for trees growing on the Niagara escarpment¹. They must be some of the slowest-growing woody plants on Earth. All the trees were grossly deformed as well as stunted. A 1.5-metre-tall, 1,140-year-old *J. phoenicea* tree, for example, from the Verdon Gorge, France, had a radius of nearly 8 cm and an annual radial growth rate slightly more than 0.06 mm. Its axial morphology was typical of cliff trees in exhibiting partial cambial mortality and an inverted morphology. All core samples from cliff-face trees showed radial growth rates that were the same in the sapling and the mature stages of development.

The ages of the trees may indicate the ages and growth rates of the entire plant communities on the cliffs. Cliffs across the world may support ancient, slow-growing, open woodland communities that have escaped major human disturbance, even when they are situated close to agricultural and industrial activity, which has destroyed or altered most other natural habitats⁸⁻¹⁰.

D. W. Larson*, **U. Matthes***, **J. A. Gerrath***, **J. M. Gerrath†**, **J. C. Nekola‡**, **G. L. Walker§**, **S. Porembski¶**, **A. Charlton#**, **N. W. K. Larson***

*Cliff Ecology Research Group, University of Guelph, Guelph, Ontario, N1G 2W1, Canada
e-mail: dwlarson@uoguelph.ca

†Department of Biology,

University of Northern Iowa, Cedar Falls, Iowa 50614, USA

‡Department of Natural and Applied Sciences, University of Wisconsin, Green Bay, Wisconsin 54311, USA

§Department of Biology, Appalachian State University, Boone, North Carolina 28608, USA

¶Botanisches Institut, University of Bonn, 53115 Bonn, Germany

#Department of Biology, University of Manchester, Manchester M13 9PT, UK

1. Larson, D. W. & Kelly, P. E. *Can. J. Bot.* **69**, 1628–1636 (1991).
2. Kelly, P. E. & Larson, D. W. *J. Ecol.* **85**, 467–478 (1997).
3. Kelly, P. E., Cook, E. R. & Larson, D. W. *Can. J. Forest. Res.* **24**, 1049–1057 (1994).
4. Zar, J. H. *Biostatistical Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1974).
5. Conover, W. J. *Practical Non-parametric Statistics* 2nd edn (Wiley, New York, 1971).
6. Hett, J. M. & Loucks, O. L. *J. Ecol.* **64**, 1029–1044 (1976).
7. Kelly, P. E., Cook, E. R. & Larson, D. W. *Int. J. Plant Sci.* **153**, 117–127 (1992).
8. Vitousek, P. M. *Ecology* **75**, 1861–1876 (1994).
9. Spencer, J. W. & Kirby, K. J. *Biol. Conserv.* **62**, 77–93 (1992).
10. Peterken, G. F. *Natural Woodland* (Cambridge Univ. Press, 1996).

Molecular basis of triclosan activity

Triclosan (5-chloro-2-(2,4-dichlorophenoxy) phenol) has been used for more than 30 years as a general antibacterial and antifungal agent, and is found in formulations as diverse as toothpastes, cosmetics, antiseptic soaps, carpets, plastic kitchenware and toys. It has recently been suggested that triclosan blocks lipid biosynthesis by specifically inhibiting the enzyme enoyl-acyl carrier protein reductase (ENR)¹. We have carried out a structural analysis and inhibition experiments on a complex of ENR from the bacterium *Escherichia coli* with triclosan and NAD⁺. We find that triclosan acts as a site-directed, very potent inhibitor of the enzyme by mimicking its natural substrate.

ENR catalyses the final, regulatory step in the fatty-acid synthase cycle: the reduction of a carbon-carbon double bond in an enoyl moiety that is covalently linked to an acyl carrier protein. ENR has been identified as a target for the diazaborines², a family of antibacterial agents, and many triclosan-resistant strains of *E. coli* have been found to be resistant to diazaborines¹.

The structure of the *E. coli* ENR complexed with NAD⁺ and triclosan (triclosan was provided by Ciba Speciality Chemicals and Coalite Chemicals) was determined to 2.2 Å resolution. The electron density map of the complex is of high quality (Fig. 1a) and reveals the mode of binding of the triclosan adjacent to the nicotinamide ring of the nucleotide cofactor in the enzyme's active site. The phenol ring of triclosan forms a face-to-face interaction with the

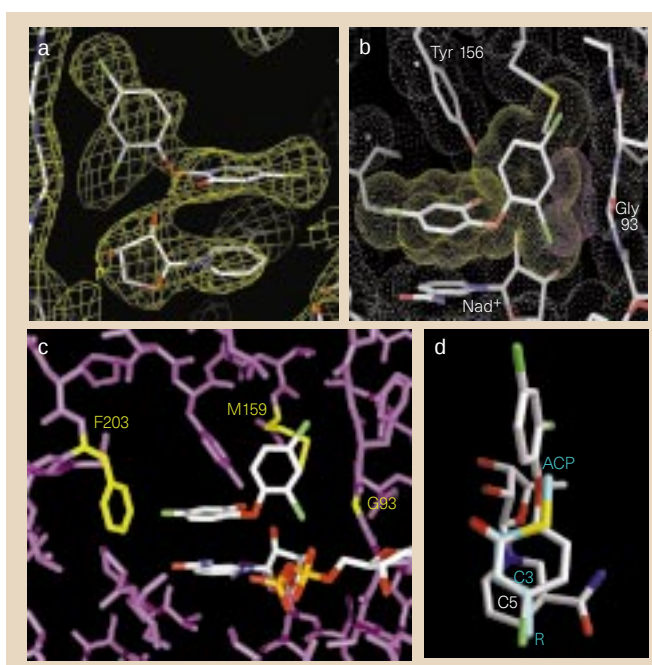
nicotinamide ring, allowing extensive π - π stacking interactions (Fig. 1b). Additional van der Waals contacts are made by both rings of the triclosan with residues lining the active site and the substrate-binding pocket of the enzyme and with the nucleotide cofactor. Hydrogen bonds are formed by the phenolic hydroxyl group of triclosan with the 2' OH of the nicotinamide ribose of the nucleotide and with the phenolic oxygen of tyrosine 156, which is believed to function as a proton donor during the catalytic cycle of ENR³.

Triclosan resistance in *E. coli* K12 is conferred by single amino-acid substitutions in ENR: G93V, M159T and F203L (ref. 1). Analysis of the structure of the *E. coli* ENR-NAD⁺-triclosan complex reveals that the three substitutions affect residues that have direct contacts with the triclosan (Fig. 1c). The most dramatic substitution occurs at position 93 of the enzyme and leads to the highest level of resistance, a roughly 100-fold increase in the minimal inhibitory concentration. The G93V substitution would cause adverse steric contacts between the valine side chain at this position and atoms in the 2,4-dichlorophenoxy ring of triclosan (Fig. 1b). As noted previously¹, the amino-acid substitutions in *E. coli* ENR that confer triclosan resistance are located adjacent to the enzyme's diazaborine binding site⁴. In particular, the G93V replacement, which leads to a dramatic weakening of triclosan binding, is also the site of the G93A/S/C/V substitutions, which greatly reduce diazaborine binding to *E. coli* ENR, again through the introduction of adverse steric contacts⁵.

The catalytic mechanism of ENR is thought to proceed by hydride transfer from the C4 position of NADH to the C3 position of an enoyl moiety that is covalently linked to the phosphopantetheine arm of the incoming acyl carrier protein. This results in the substrate being rearranged to produce an enolate anion intermediate before subsequent proton donation and product formation³. Analysis of a model for the binding of the enoyl moiety to ENR and the binding of triclosan indicates that the structure of the enolate anion intermediate of the substrate is closely mimicked by part of the phenolic ring of the triclosan (Fig. 1d). The enoyl C3 carbon, to which the hydride is transferred during catalysis, and the enolate oxygen appear to bind at the same sites as the carbon at position 5 of the phenolic ring and the phenolic oxygen of triclosan, respectively.

The similarity in the structure of triclosan to the enzyme's natural, tightly binding substrate (the apparent Michaelis constant, K_M , for crotonyl ACP is 22 μ M)⁶ and the complementary nature of the triclosan-binding site indicate that the inhibitor binds tightly to the enzyme. As

Figure 1 Triclosan bound to *E. coli* ENR. a, Fourier map at 2.2 Å resolution (1σ contour; coefficients $2|F_o| - |F_c|$) with the refined structure, from co-crystals of an ENR/NAD⁺/triclosan complex that are isomorphous with those formed by an ENR/NAD⁺/diazaborine complex⁴. Data were collected with 78% completeness to 2.2 Å and $R_{\text{merge}} = 0.057$. Triclosan was located by difference Fourier analysis and the structure was refined to an R -factor of 0.241 (data in the range 10–2.2 Å; free R -factor, 0.286). Generated using the program O (ref. 8). b, Van der Waals surfaces of the enzyme, NAD⁺ and triclosan (O, red; N, blue; S, yellow; C, white; P, orange; Cl, green) calculated at 95% radii, shown as white dots for ENR and NAD⁺ and yellow dots for triclosan. A mutation of glycine 93 to valine, shown as magenta dots at 75% radii for the side chain of valine, shows how severe steric clashes would result from overlap of the triclosan and valine surfaces. Produced using SYBYL v5.41 (Tripos). c, Residues G93, M159 and F203 (bright yellow) lead to triclosan resistance; the rest of the enzyme is purple, and triclosan and NAD⁺ moieties are coloured as in b. d, Superposition of bound triclosan and a model for the enolate anion substrate intermediate showing how triclosan mimics the substrate by similar binding of part of its phenolic ring and the proposed structure of the bound substrate. The NAD⁺ nicotinamide ring and associated ribose are at the back and coloured, like triclosan, as in b; the substrate's carbon atoms are cyan. Predicted locations of the phosphopantetheine arm (ACP) and growing acyl chain (R) of the substrate are marked. c, d, Produced using Midas⁹.



enzyme activity progressively decreased with time, in further *in vitro* assays we preincubated the enzyme with NAD⁺, which gave 50% inhibition of initial velocity for 120 nM triclosan. Reaction mixture (100 μ l), containing 10 mM sodium phosphate, pH 8.0, 0.14 mM NADH, 0.14 mM NAD⁺ and 200 nM wild-type *E. coli* ENR enzyme, was preincubated with triclosan dissolved in 50% dimethylsulphoxide (final 0.5%) for 2 min before crotonyl-CoA (K_M 2.7 mM)⁶ was added, to a final concentration of 2.5 mM. This concentration of triclosan is about half the enzyme concentration used in the assay mixture, indicating that triclosan is a very potent inhibitor.

This finding is consistent with *in vivo* inhibition studies¹ showing that triclosan is active at very low concentrations. In contrast, other studies have suggested rather weaker binding, with a half-maximal inhibitory concentration (IC_{50}) of 2.0 μ M for inhibition of ENR by triclosan⁷. But these experiments did not involve a pre-incubation step with NAD⁺, and our results show a more marked inhibition in the presence of the oxidized cofactor.

Our structural and kinetic data provide a view of the tight binding between an enzyme that is potentially important as a chemotherapeutic target and a very potent and specific inhibitor, which is commonly

used in the home and the workplace. Our findings support the view that triclosan acts mainly by specifically inhibiting fatty-acid biosynthesis.

Colin W. Levy*, **Anna Roujeinikova***, **Svetlana Sedelnikova***, **Patrick J. Baker***, **Antoine R. Stuitje†**, **Antoni R. Slabas‡**, **David W. Rice***, **John B. Rafferty***

*Krebs Institute for Biomolecular Research, Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, UK

†Department of Genetics, Institute of Molecular Biological Studies (IMBW), Vrije Universiteit, Biocenter Amsterdam, 1081 HV Amsterdam, The Netherlands

‡Department of Biological Sciences, University of Durham, Durham DH1 3LE, UK

e-mail: d.rice@sheffield.ac.uk

- McMurry, L. M., Oethinger, M. & Levy, S. B. *Nature* **394**, 531–532 (1998).
- Turnowsky, F., Fuchs, K., Jeschek, C. & Hogenauer, G. *J. Bacteriol.* **171**, 6555–6565 (1989).
- Baldock, C., Rafferty, J. B., Stuitje, A. R., Slabas, A. R. & Rice, D. W. *J. Mol. Biol.* **284**, 1529–1546 (1998).
- Baldock, C. et al. *Science* **274**, 2107–2110 (1996).
- Boer, G.-J. et al. *Mol. Microbiol.* **31**, 443–451 (1999).
- Bergler, H. et al. *J. Biol. Chem.* **269**, 5493–5496 (1994).
- Heath, R. J., Yu, Y.-T., Shapiro, M. A., Olson, E. & Rock, C. O. *J. Biol. Chem.* **273**, 30316–30320 (1998).
- Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. *J. Mol. Graph.* **6**, 13–27 (1988).

Evolutionists red in tooth and claw

The Darwin Wars: How Stupid Genes Became Selfish Gods

by Andrew Brown

Simon & Schuster: 1999. 241 pp. £12.99, \$20

David L. Hull

Dust-jackets are frequently adorned by quotations from famous people praising the book. At first glance, Andrew Brown's *The Darwin Wars* is no exception. Pithy quotations from Steve Jones, Richard Dawkins, John Maynard Smith, Stephen Jay Gould and Daniel Dennett. Who could ask for more? However, on closer inspection these quotations turn out not to be about Brown's book at all, but quotations that Brown uses in his book. Only Dennett's blurb refers to one of Brown's own publications: "What a sleazy bit of trash journalism."

The Darwin Wars chronicles the battles over the past 25 years among warring factions in evolutionary biology—the Dawkinsians versus the Gouldians, as Brown dichotomizes the conflict. In his narrative, Brown interweaves interpersonal relationships with issues of substance. For example, E. O. Wilson was especially hurt when colleagues whom he considered long-time friends viciously attacked his book on socio-biology, generating years of nasty exchanges in *The New York Review of Books*, the form of fisticuffs preferred by American academics.

Brown also details the issues that separated the combatants. For example, he discusses at length David Haig's work on the warfare between a mother and her fetus, Elaine Morgan's aquatic ape theory, Napoleon Chagnon's warring Yanomamö, Margo Wilson's explanation of why children are more often murdered by their mother's new boyfriend than by their biological father, and John Tooby and Leda Cosmides's arguments against the mind being a general-purpose tool. He devotes an entire chapter to the adaptive advantages of religious beliefs. In each case, Brown reiterates the familiar message that adaptationist claims are extremely difficult to check. If we accept evolutionary theory, as everyone in this book does, then some adaptationist claims must be true. The problem is deciding which ones.

Brown is a freelance journalist, a precarious way of making a living. As a journalist he must write well, at least well enough to keep his readers turning the pages of his books. From my perspective, Brown's first chapter contains his most successful prose. He relates a tragic story of the intellectual and personal relationships between three highly original evolutionary biologists, two well known and one obscure. In the late 1950s, William Hamilton studied with John Maynard Smith at University College London. Maynard



Smith hardly noticed Hamilton. A decade later, both men independently came to realize how original the ideas of an amateur biologist named George Price were. Price had worked out a mathematical argument showing that altruistic behaviour is possible but that it is not in the least bit noble. According to Brown, this proof sent Price into a deep depression that was lifted only by a religious revelation that was visited upon him just north of the BBC Broadcasting House. Thereafter, Price tried to combine his work on the evolutionary process with ministering to the down-and-outs in London's back streets. It was not long before he had joined his charges as he descended into near madness. In the winter of 1974 Price killed himself in a squat near Euston Station, driven to suicide, as Brown sees it, by his work on the evolution of altruism and selfishness.

In many respects, *The Darwin Wars* is a remake of the tired old nature-nurture scenario. Back in the 1930s the environment was in the driver's seat—no matter how horrific the crime, the response among the literate public tended to be reform and counselling. It was not his fault: his environment made him do it. More recently, genes have caught the imaginations of journalists and popularizers. It was not his fault: his genes made him do it. Today all sides agree that both nature and nurture are important in making us what we are, but the Dawkinsians emphasize the role of genes while the Gouldians emphasize culture. Hence, on the surface, the issue in *The Darwin Wars* is the relative mix of genes and environment.

But the deeper issue is not nature-nurture at all, but free will. Can our environment and genes gang up on us so that we are forced to follow their dictates? All sides answer that they can't, but exactly how such resistance is possible is not easy to say. Brown reveals

himself as a closet Gouldian when he explains Price's madness in terms of his environment, not his genes. Although Brown acknowledges that "one cannot say exactly what drove Price mad", he attributes his mental disintegration and suicide to his failure to come to terms with Hamilton's equations. The primary cause might just as well have been genes.

Brown makes a mistake or two. For example, he identifies "being a woman" with "possessing a Y chromosome". In relating the early history of Mormonism in the United States, he mentions how Joseph Smith founded Nauvoo in Illinois on the banks of the Mississippi river—I know, because I was born only a few miles from Nauvoo, and some of my ancestors were part of the "lynch mob" that shot Smith.

One of the strengths of Brown's exposition is that he does not pretend to be a disembodied intellect hovering above the fray. In similar circumstances he is likely to behave in similar ways. One thing that does come through is that the scientists Brown discusses really care about their science. The combatants "write from conviction, not for hire". That is why their disputes become so acerbic. If you attack mine, you attack me. Is Brown's book one more "sleazy bit of trash journalism"? I don't think so, but my judgement might be coloured by the fact that Brown treats my work very gently and relates no painful stories about me and my professional friends. □

David L. Hull is in the Department of Philosophy, Northwestern University, Evanston, Illinois 60208, USA.

On from Babylon

The History and Practice of Ancient Astronomy

by James Evans

Oxford University Press: 1998. 478 pp.

\$65, £49.50

J. D. North

James Evans is a historian of astronomy who teaches both physics and the history of science at the University of Puget Sound in Washington state. Such biographical details offer a clue to his chosen genre, which is not only unusual but, sadly, unfashionable. Evans approaches the history of ancient astronomy determined to enter the minds of those whose work he discusses, and in this, within his chosen limits, he is highly successful. His is a hands-on approach eminently fitted to the needs of students who want to know not only what historians have said



A Greek heaven: the ancient Greeks believed that the heavens were spherical, and that the Earth lay at its centre. Atlas's hand supports this oldest known celestial sphere, the Farnese globe.

about ancient astronomy, but also how astronomy was actually done in the past.

The book is copiously illustrated with helpful geometrical figures, and supplemented here and there with a generous touch of simple mathematics, but it will be accessible only to readers above a certain level of numeracy. Evans deals with material ranging from the Babylonian arithmetical analyses of planetary motion to Greek epicycles, and ultimately to Keplerian ellipses, presenting the reader with material from the very core of the exact sciences over the past three millennia.

Most of his story is not new, but what is new is his way of telling it. Whether this book can be described as history is a concern to which I imagine its author is fairly indifferent. There is history running quietly through its pages, but it takes second place and is regularly interrupted with problems and exercises for the reader that are designed to give a feeling for the practical art of astronomers from the past. However, Evans' attention to past practice does not mean that he avoids either the literary or the philosophical traditions, for both have a foot in popular, home-

spun, practical astronomy, and so earn a small place here.

The earliest activity meriting the name astronomy needed no instrumental aid. The classic case of heliacal risings and settings is a good example. Yet instruments naturally enter the story at an early stage, and the book's greatest strength is in the clarity with which Evans explains how they work. Much can be learned of solar motions even from the humble gnomon, which is first on the list. The usefulness of the illustrations, line drawings, tables and reproductions of old prints cannot be exaggerated. You won't find every history of ancient astronomy illustrating the principles of an astrolabe plate with an example for Mexico City.

Students who learn the elements of spherical astronomy from this book will certainly have a firm grasp of what is involved. However, Evans' account will give them a set of conceptual devices — the celestial sphere and associated geometrical elements — that stem from the Greek tradition, and which was not shared by the Babylonians or early Egyptians, for instance. Evans is aware of the danger that, by structuring his material

around topics such as the celestial sphere, calendars and solar theory, he might obscure the historical picture. The attentive reader might find his remedy tucked away in a corner near the beginning of the book. Thus, he advises: "The reader is invited to return to this survey to see how a particular writer or topic fits into the broader picture." Whether this rings true to the real world of student psychology is a moot point, for this bumper book is not for the faint-hearted. A student who works steadfastly through it from beginning to end will have read something over a quarter of a million words. The extra stamina required to superimpose the "broader picture" will seem a small investment to the dedicated reader who has covered this ground — always supposing a historical dimension is wanted. But I know from experience that many science students want no more than Evans offers them here. Should they be force-fed?

And what of other potential readers, the intelligent dabbler, for instance? One of the beauties of the book is that it can be browsed non-consecutively. In fact, some sections could easily be used as self-contained course material — the chapter on the calendar, for instance, would provide a better survey for budding historians with chronological tendencies than much of the recent semi-popular literature on that subject.

There is a broadly historical pattern within each chapter that tends to begin with Babylon, or occasionally ancient Egypt, pays most attention to ancient Greece, and then moves quickly through mediaeval Islam and Christian Europe, with some emphasis on the period immediately before and after Copernicus. In a sense, this book is the story of Copernicanism as a recasting of ancient, and in particular Ptolemaic, astronomy.

There is a distinct undertone of scientific progressionism throughout, despite the occasional sceptical remark about the obsessive search for precursors. The question of Copernicus' reliance on his predecessors is sensitively handled, and lacks the hyperbole so often associated with it in recent decades. Evans touches lightly on this and on various other modern controversies over priority, notably that concerning Ptolemy's putative plagiarism of Hipparchus, an idea vigorously promoted by the late Robert R. Newton in the 1970s. In this last case, Evans adds his own extended analysis, which is both simple and convincing, tacitly defending Ptolemy by revealing flaws in his critics' arguments. This is an excellent example of Evans' talent for dispassionately evaluating the evidence, which pervades his text. He offers a no-nonsense account that breathes life into the astronomical past while remaining true to his subject.

His is not the whole truth, and one might occasionally question his choice of emphasis, but it would be hard to find an

elementary account of the subject that is as accurate or as comprehensive as his. I suspect that most students who emerge victorious from this tough assignment will tacitly conclude that they share a common scientific psychology, even motivation, with those whose methods they can now so well understand. In this they are likely to be mistaken, but the fact that the inference is partly right is what makes Evans' book so educationally important. □

J. D. North is at the Philosophical Institute, University of Groningen, A-Weg 30, 9718 CW Groningen, The Netherlands.

The laying on of genes

The Development of Human Gene Therapy: Monograph 36

edited by Theodore Friedmann
Cold Spring Harbor Laboratory Press: 1999.
729 pp. \$134

Karol Sikora

The transfer of specific genetic material to treat disease has been one of the most exciting potential applications of modern genetics. Given the complexity of the task, it is perhaps not surprising that progress has been slow in terms of clinical gain. The main stumbling block right from the start has been in getting DNA to the right place in a form that permits its expression at the right time in enough cells. The problem of effective *in vivo* gene targeting still bedevils the field a decade after the first trials began.

The collection of reviews in *The Development of Human Gene Therapy* begins with a historical overview going back to Archibald Garrod, the senior physician at St Bartholomew's Hospital, London, who, around the turn of the century, discovered what he termed the "inborn errors in metabolism". It covers the abortive and potentially unethical beginnings of human globin-gene transfer for thalassaemia in the early 1970s, when globin genes were introduced into the bone marrow cells of two patients. Gene expression was unlikely, and no ethical approval had been obtained. The first successes came with cystic fibrosis, adenosine-deaminase deficiency and cancer. Success, of course, is only relative. Real success equates with clinical benefit in terms of improved quality of life or increased survival, and this has so far proved elusive.

The chapters on the major vehicles for gene delivery cover retroviruses, adenoviruses, herpes viruses, lentiviruses, a range of more unusual vectors and, of course, physical delivery systems. Much of the information is dreadfully old and can be found in greater detail elsewhere. Although the major targets for clinical exploitation

are discussed, it seems strange to omit the remarkable clinical developments in cystic fibrosis, ischaemic heart disease and the rare single-gene disorders. Some clinical possibilities that are now the focus of much current activity, such as arthritis, retinopathies, collagen disorders, asthma, diabetes and obesity, are not even mentioned.

Highlights of this large volume include an excellent chapter on naked DNA injection into various tissues. The fact that this leads to gene expression at all is remarkably encouraging for the future. Another chapter considers targeted gene repair in mammalian cells using chimaeric oligonucleotides. Although a long way from the clinic, this represents the beginnings of genetic surgery. The short review on stem-cell transplantation is concise, yet informative, while the ethics of the subject are well reviewed in a historical context.

There are several excellent short books reviewing gene therapy. Unfortunately, this is not one of them; its content is outdated and the artwork incredibly poor, and, a major criticism, despite the encyclopaedic coverage of the field, the material comes almost exclusively from US-based authors. Human-gene therapy is now an international endeavour, with clinical protocols in place in more than 25 countries. True, the Americans got there first, but please, the rest of the world does exist. □

Karol Sikora is at the World Health Organization Cancer Programme, International Agency for Research on Cancer, 69372 Lyon, France.

Genetics of another animal

The Zebrafish: Biology (*Methods in Cell Biology* Vol. 59)

edited by H. William Detrich III,
Monte Westerfield and Leonard I. Zon
Academic: 1998. \$64.95

An explosive laboratory

Island Biogeography: Ecology, Evolution and Conservation

by Robert J. Whittaker
Oxford University Press: 1998. 285 pp.
£19.99, \$29.95 (pbk)

Ilkka Hanski

The island of Krakatau in the Sunda strait was reduced to one-third of its size by violent volcanic eruptions on 27 August 1883. Remains of unlucky life were buried under 60–80 metres of ash, but, luckily for science, the biologists of the time were quick to grasp the unique opportunity for observing recovery.

Within a few years, coastal plant communities were well established. The interior was first colonized by ferns and grasses, to be

replaced by forest within 40 years. So far, more than 300 plant species have been recorded from Krakatau, although there are fewer species there today — many species recorded in the past have become extinct. Animals present much the same picture. For instance, the cumulative number of bird species is nearly 50, whereas the present bird community has some 30 species.

Robert Whittaker's *Island Biogeography* delivers a sweeping lesson in island ecology enhanced by the author's long-term involvement in the research on Krakatau. Intriguingly, bird colonization of Krakatau also featured as the opening example in the trail-blazing island-ecology book of the century, Robert H. MacArthur and E. O. Wilson's *The Theory of Island Biogeography* (Princeton University Press). Back in 1967, these authors concluded that the number of bird species had reached an equilibrium by the 1930s, although further extinctions and colonizations of individual species were to be expected. In the graph shown by Whittaker, the number of bird species has increased only slightly since the 1930s. Some extinctions and colonizations are attributed to successional changes in the island's vegetation rather than to stochastic factors.

One might think that the latest figures endorse MacArthur and Wilson's prediction, but Whittaker takes the opposite view. In fact, by his judgement the MacArthur–Wilson theory is essentially dead, as islands are ceaselessly bombarded by disturbances — though Krakatau may be an extreme example — and there is therefore no time for an equilibrium to become established. The 'island theory' is replaced by 'disturbed island ecology', without the word 'theory' in the latter label.

Whittaker's conclusion is curious, since it is akin to asserting that density-dependent population regulation is unimportant because population sizes are perturbed by varying environmental conditions. True, an equilibrium in the sense of a constant number of individuals in a population, or of species on an island, is unlikely, but equilibrium in the sense of a characteristic distribution of numbers in the course of time may well occur. What matters are the processes.

Island Biogeography covers the full range of topics, from the geography and geology of islands, their speciation and evolution, population ecology and community assembly, to the application of island theory to conservation. Although islands in the sea account for only 3% of Earth's land area, and individual islands have impoverished communities, the pooled number of species on islands is great. For example, around 15% of all known species of birds and plants occur on islands. Islands are ecological and evolutionary laboratories. Whittaker describes cases of

adaptive radiation on islands, patterns in the adaptation to island life, assembly of island communities — and sad stories of the demise of unique island life brought about by humans and human companions from the mainlands.

As islands have a limited area, island populations have a limited size. One fascinating topic that Whittaker does not discuss — probably for lack of primary research — is the question of the survival and evolution of small populations on islands. Have some island populations the means of surviving in spite of small population size, and if so, what are these means?

Island Biogeography will satisfy those looking for a comprehensive text on island ecology, but it offers no surprises. Whittaker's 'disturbed island ecology' does not really take off. Nature is complex, but merely exposing complexity is not enough — I will keep wiping the dust off my copy of *The Theory of Island Biogeography*. □

Ilkka Hanski is in the Department of Ecology and Systematics, PO Box 17 (Arkadiankatu 7), FIN-00014, University of Helsinki, Finland.

New Journals

This year, *Nature's* annual New Journals review supplement will appear in the issue of 2 September. Publishers and learned societies are invited to submit journals for review, as well as details of any eligible electronic journals, taking note of the following criteria:

- Journals must have first appeared during or after June 1997 and issued at least four separate numbers by the end of May 1999.
- Journals covering any aspect of science are eligible, although those dealing with clinical medicine and pure mathematics are excluded, as are newsletters and publications of abstracts.
- Frequency of publication must be at least three times a year.
- The main language must be English.
- Deadline for submission is 28 May.

Please send at least four different issues (the first, the most recent and any two others) of each eligible title, together with full details of subscription rates, to: Isobel Flanagan, *Nature*, Porters South, Crinan Street, London N1 9XW, UK. Tel: +44 (0)171 843 4542. e-mail: i.flanagan@nature.com

Science in culture

Glazes from deep time

The ceramic imagery of Joan Lederman
Martin Kemp

Ceramic vessels have borne profound meaning for many cultures and civilizations, and are not merely functional objects that, at their best, aspire to a modest kind of beauty. In modern Western culture a pot tends to be a pot, and the person who makes it a potter, a craftsperson. Joan Lederman calls herself a "potter and imagineer". The extended title is well deserved; she is creating a form of ceramic art that resounds thrillingly across time and space.

On 16 July 1996, Chris Griner, an able-bodied seaman in the palaeoclimate research team at the Woods Hole Oceanographic Institution on Cape Cod, brought to Lederman's "The Soft Earth"

studio a bucket of sediment raised from the

Atlantic floor east of Virginia (latitude 32° 21' N, longitude 68° 50' W) at a depth of 4,500 m — some two miles. Fired in her kiln at 1,100 °C, lava-like, it congealed.

Three months later, Lederman experimented with the mud as a glaze for stoneware vessels. Thinned gently with water but otherwise unmodified and applied evenly over the clay bodies, the fired sediments wondrously metamorphosed from a uniform layer into modelled patterns and beguiling arrays of colour.

A series of experiments followed in a new, hotter kiln using mud not only from the waters around the United States but also from such far-flung sources as the Arabian and Indian Oceans, the latter found by carbon dating to be 35–40 million years old. The example illustrated, which used sediments brought back to Woods Hole by the research vessels *Knorr* and *Oceanus*, shows how the resulting glazes "seem to have a life of their own — like a fingerprint". While the "glazes that humans have concocted are often exquisite ... they have a different voice. They are more contrived from nature than they are released by nature."

Lederman sees herself as "a channel for what the sea muds might do ... I adjust the form, the application, the thickness, the claybody, the firing, the juxtaposition to other glazes, to make what could happen actually emerge."

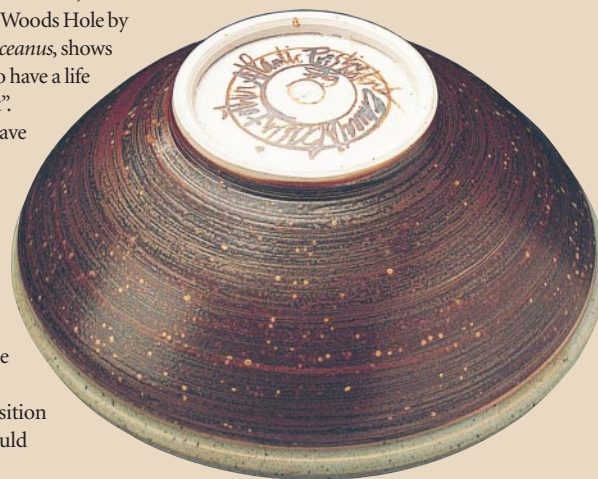
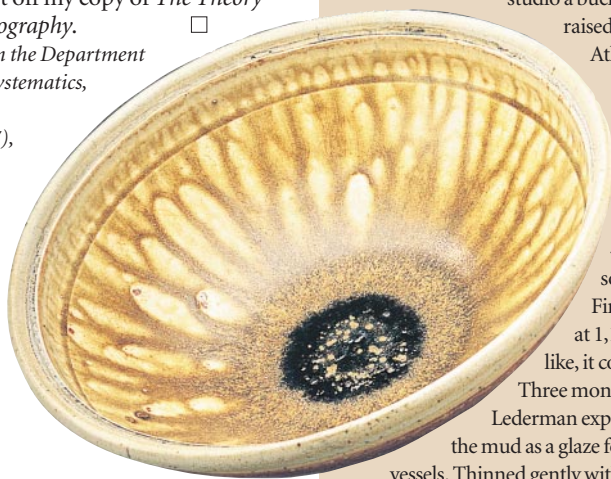
What emerges, most notably with sediments that are dense in foraminifera, as on the inside of the bowl, are beautiful variations on dendrite formations, seemingly rising and ramifying, brought, from within the inferno of Lederman's kiln, to new life after aeons of remote stillness in the depths of the sea. On one level, we can see the results as beguiling examples of the potter's art, but on another they carry an extraordinary *frisson*. To cradle one of Lederman's vessels in one's hands is an evocative, even eerie experience, akin to holding a fragment of meteorite or peering at a sample of Moon rock.

Visually, the emergent patterns resonate with the dendritic forms that comprise one of the recurrent organizational configurations in nature, ranging from roots of nerve cells to river deltas. And they can now be modelled on computer through diffusion-limited aggregation, in which particles in Brownian motion adhere successively around a fixed particle to generate fern-like radiations.

Lederman is knowingly leaving the field for perception and interpretation open, creating "holistic experiences that call people to the learning that is alive for them". But in her collaboration with the ancient muds that hold the secrets of palaeontological climates, she is actively inducing us to share her sense of wonder at the way earth, fire and water — the very stuff of the potter's art — speak of elemental forces that have repeatedly attracted human creators across distant times and spaces. □

Joan Lederman's ceramics can be seen in the New England Aquarium, Boston, Massachusetts 02110, USA, and at <http://www.arts-cape.com/softearth/>. Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

Inner and outer views of the North Atlantic Deep Sea Bowl, glazed with sediments recovered by the research vessels *Knorr* and *Oceanus*, July 1998.



JOAN LEDERMAN

The afterglow, redshift and extreme energetics of the γ -ray burst of 23 January 1999

S. R. Kulkarni¹, S. G. Djorgovski¹, S. C. Odewahn¹, J. S. Bloom¹, R. R. Gal¹, C. D. Koresko¹, F. A. Harrison¹, L. M. Lubin¹, L. Armus², R. Sari³, G. D. Illingworth⁴, D. D. Kelson⁵, D. K. Magee⁴, P. G. van Dokkum⁶, D. A. Frail⁷, J. S. Mulchaey⁸, M. A. Malkan⁹, I. S. McClean⁹, H. I. Teplitz¹⁰, D. Koerner¹¹, D. Kirkpatrick², N. Kobayashi¹², I.-A. Yadigaroglu¹³, J. Halpern¹³, T. Piran¹³, R. W. Goodrich¹⁴, F. H. Chaffee¹⁴, M. Feroci¹⁵ & E. Costa¹⁵

Long-lived emission, known as afterglow, has now been detected from about a dozen γ -ray bursts. Distance determinations place the bursts at cosmological distances, with redshifts, z , ranging from ~ 1 to 3. The energy required to produce these bright γ -ray flashes is enormous: up to $\sim 10^{53}$ erg, or 10 per cent of the rest-mass energy of a neutron star, if the emission is isotropic. Here we present optical and near-infrared observations of the afterglow of GRB990123, and we determine a redshift of $z \geq 1.6$. This is to date the brightest γ -ray burst with a well-localized position and if the γ -rays were emitted isotropically, the energy release exceeds the rest-mass energy of a neutron star, so challenging current theoretical models of the sources. We argue, however, that our data may provide evidence of beamed (rather than isotropic) radiation, thereby reducing the total energy released to a level where stellar-death models are still tenable.

Almost 30 years after their discovery¹, the distance scale to γ -ray bursts (GRBs) was finally unambiguously determined just two years ago². Thanks to precise and prompt GRB localizations by the Italian-Dutch satellite, BeppoSAX³, and the All Sky Monitor⁴ on board the Rossi X-ray Timing Explorer, afterglow emission spanning the wavelength range from X-ray to radio has now been detected in about a dozen events^{5–7}. This has provided redshift measurements for four bursts: one through the detection of optical absorption lines², and three by identifying emission features in the host galaxies^{8–10}. It is now assumed that a substantial fraction of all GRBs occur at significant cosmological distances ($z \approx 1–3$).

The large inferred distances imply staggering energetics. In particular, our measured redshift of $z = 3.42$ for GRB971214 (ref. 8) led us to an inferred isotropic energy loss of 3×10^{53} erg, or $0.1M_{\odot}c^2$ (where $M_{\odot}$ is the solar mass, and c is the velocity of light). At that time, this result was seen to challenge most proposed GRB models. The energy estimate can be reduced by invoking non-spherical emitting surface geometry, such as jets. But to date, there is no firm observational evidence for jets in GRBs. Indeed, on statistical grounds, Grindlay¹¹ and Greiner *et al.*¹² constrain the “beaming factor” (the reduction in inferred energy release relative to that for the spherical case) to be no less than 0.1.

BeppoSAX ushered in 1999 with the discovery of GRB990123¹³, the brightest GRB seen by BeppoSAX to date. It is in the top 0.3% of all bursts, if ranked by the observed fluence¹⁴. If the inferred energetics of GRB971214 strained theoretical models, GRB990123, with a minimum redshift of 1.6 and an inferred isotropic energy release of $1.9M_{\odot}c^2$ takes them to the breaking point. Here we present optical and infrared observations of

GRB990123, which both establish the distance to this luminous event, and—through the observed break in the decay—may provide the first observational evidence yet of beaming in a GRB.

The optical and the infrared afterglow

Much of our information about the parameters of GRB explosions (energy, ambient gas density, dynamics) has been obtained from radio and optical afterglow studies. These returns have motivated observers to carry out long-term multi-wavelength studies of the afterglow.

We initiated an optical and infrared follow-up program for GRB990123 on 23.577 January 1999 UT, 3.7 h after the event, in response to the BeppoSAX preliminary localization¹⁵. We used a charge-coupled device (CCD) at the Palomar 60-inch telescope to image the field of the BeppoSAX localization. We identified a new source by comparing the first CCD image with an image of the field obtained from the Digital Palomar Observatory Sky Survey II (DPOSS), and concluded that this was the optical afterglow of GRB990123. We promptly reported the discovery to the GRB Coordinates Network¹⁶ (GCN) and the Central Bureau of Astronomical Telegrams. The optical source lies within the tighter BeppoSAX localization¹³ obtained a few hours later (see Fig. 1).

Continued observations at Palomar, Keck, MDM and other observatories showed that the source exhibited the characteristic fading behaviour seen in other GRB optical afterglows. The association of the optical transient (OT) with GRB990123 is therefore secure. Table 1 summarizes the photometric observations of the OT. Included in the table is a single observation¹⁷ obtained from the Hubble Space Telescope (HST); the analysis of the HST data are reported elsewhere¹⁸. Figure 2 shows the OT light curve in various bands (primarily Gunn-r and K) using the photometric data shown in Table 1.

A considerable body of literature^{19–22} has been developed to explain the afterglow phenomenon. Briefly, afterglow can be understood as arising from synchrotron emission from particles shocked by the explosive debris sweeping up ambient medium. Assuming that the electrons behind the shock are accelerated to a power-law differential energy distribution with index $-p$, the simplest afterglow model (for example, ref. 23) predicts that the afterglow flux, $f_{\nu}(t) \propto t^{\alpha} \nu^{\beta}$ where $f_{\nu}(t)$ is the flux at frequency ν and t is the time

¹ Palomar Observatory 105-24, Caltech, Pasadena, California 91125, USA; ² Infrared Processing & Analysis Center, Caltech, Pasadena, California 91125, USA; ³ Theoretical Astrophysics 130-33, Caltech, Pasadena, California 91125, USA; ⁴ Lick Observatory, University of California, Santa Cruz, California 95064, USA; ⁵ Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Rd, NW, Washington DC 20015, USA; ⁶ Kapteyn Astronomical Institute, PO Box 800, NL-9700 AV, Groningen, The Netherlands; ⁷ National Radio Astronomy Observatory, PO Box O, Socorro, New Mexico 87801, USA; ⁸ Observatories of the Carnegie Institution, 813 Santa Barbara Street, Pasadena, California 91101, USA; ⁹ Department of Physics & Astronomy, University of California, Los Angeles, California 90095-1562, USA; ¹⁰ Goddard Space Flight Center, Code 681, Greenbelt, Maryland 20771, USA; ¹¹ University of Pennsylvania, 4N14 DRL, 209 S. 33rd Street, Philadelphia, Pennsylvania 19104-6396, USA; ¹² Subaru Telescope, National Astronomical Observatory of Japan, 650 N. A'ohoku Place, Hilo, Hawaii 96720, USA; ¹³ Department of Astronomy, Columbia University, 538 W. 120th Street, New York, New York 10027, USA; ¹⁴ W. M. Keck Observatory, 65-1120 Mamalahoa Highway, Kamuela, Hawaii 96743, USA; ¹⁵ Istituto di Astrofisica Spaziale, CNR, via Fosso del Cavaliere, Roma I-00133, Italy.

measured with respect to the epoch of the GRB. Interpretation of the afterglow data in the framework of these models can, in principle, yield many interesting parameters of the GRB explosion (for example, refs 24, 25).

As the OT faded, the K-band image of 28 January showed that the host galaxy was only a fraction of an arcsecond displaced with respect to the OT. A proper analysis of the decay of the OT thus requires that we correctly assess the contribution of the host galaxy to the flux of the OT. The adopted values are $0.9 \pm 0.3 \mu\text{Jy}$ (K-band) and $0.5 \pm 0.1 \mu\text{Jy}$ (r-band); see Fig. 2 legend and ref. 18 for additional details. A discussion of the host galaxy can be found in ref. 18.

In Fig. 2 we present the light curve of the OT after subtracting the contribution from the host galaxy and correcting for Galactic extinction. No single power law can fit the r-band curve. We fit two power laws with a break at epoch t_{break} (Fig. 2). The best-fit parameters are $\alpha_{1r} = -1.10 \pm 0.03$, $t_{\text{break}} = 2.04 \pm 0.46 \text{ d}$ and $\alpha_{2r} = -1.65 \pm 0.06$. In contrast, the K-band data are well fitted by a single power law, $\alpha_K = -1.10 \pm 0.11$. We now discuss the behaviour of the afterglow in the first few days.

The value of α_{1r} and α_K are similar to those measured in other afterglows (see, for example refs 8, 26, 27). In the 2–10 keV band, the afterglow was observed by a variety of instruments on board BeppoSAX starting 6 h after the burst¹³. From the analysis of these data, $\alpha_X = -1.44 \pm 0.07$. In the framework of the afterglow models, the power-law decay index $\alpha = 3(1-p)/4$ or $\alpha = (2-3p)/4$, depending on whether the electrons are cooling on a timescale that is slower or faster than the age of the shock. A value of $p = 2.44$ is consistent with α_{1r} , α_K and α_X ; this would require that the X-ray emitting electrons are in the fast-cooling regime whereas the electrons emitting photons in the r and the K bands are in the slow-cooling regime.

As briefly summarized above, afterglow models predict that the spectrum of the OT should also be a power law and with the spectral index β having a specific relation to α . After correcting for Galactic extinction, one day after the burst we find that $\beta_{rK} = -0.8 \pm 0.1$. This value is what is expected when the electrons emitting optical and infrared photons are in the slow-cooling regime. In contrast, comparing the X-ray flux six hours after the burst¹³ with the interpolated r-band flux (from Fig. 2) we find $\beta_{rX} = -0.54$. Such

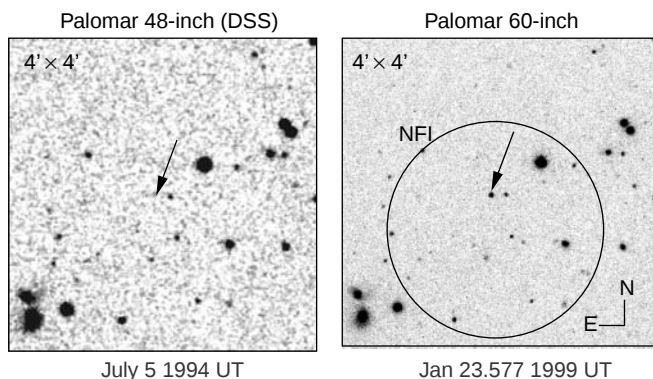


Figure 1 Discovery image of the optical counterpart of GRB990123. (Left, a $4 \text{ arcmin} \times 4 \text{ arcmin}$ view centred on the optical transient (OT) and extracted from the Digital Palomar Observatory Sky Survey II (DPOSS); the DPOSS is the digitized version of the second Palomar Observatory Sky Survey. Right, a CCD image in the Gunn-r band obtained at the Palomar 60-inch telescope. The new source, the presumed optical afterglow of GRB990123, is marked. The circle is the 50-arcsec localization of the X-ray afterglow¹³ as obtained from the Narrow Field Instrument (NFI) on board BeppoSAX. Absolute astrometry on the Palomar 60-inch discovery image was obtained by comparison of 34 objects near the optical transient with the USNO-A2.0 V2.0 Catalogue⁶⁰. The r.m.s. uncertainties of the astrometry are $0.28''$ (right ascension, RA) and $0.26''$ (declination, dec.) We find the position of the OT to be RA 15 h 25 min 30.34 s, dec. $44^\circ 45' 59.11''$ (J2000).

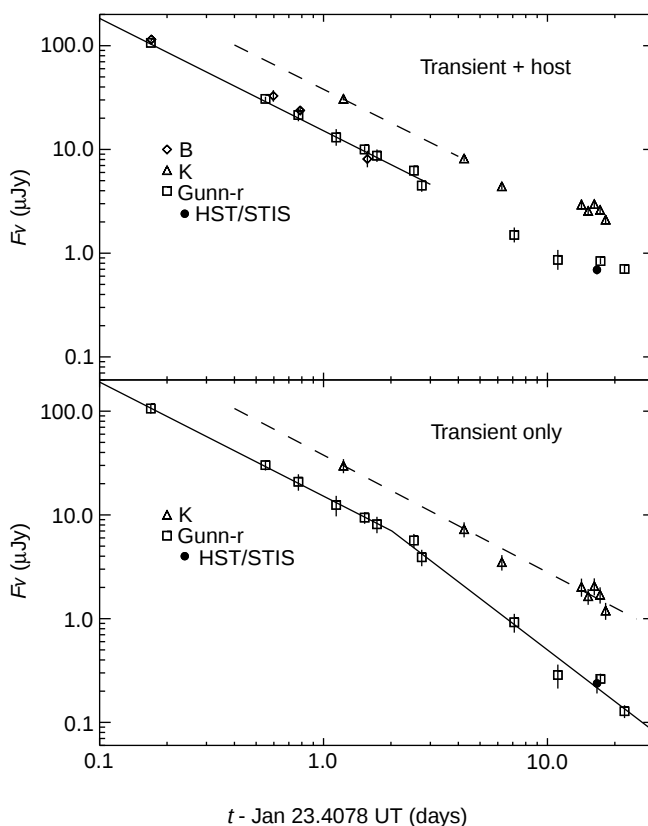


Figure 2 Optical and infrared light curves of the transient afterglow of GRB990123 from 4 hours to 23 days after the burst. The observed magnitudes from Table 1 have been corrected for Galactic extinction and converted to flux (see below). Top, the light curve of the transient + host galaxy. From ref. 61 we estimate the Galactic extinction in the direction of the optical transient (Galactic longitude 73.12° , latitude 54.64°) to be $E(B-V) = 0.01597$. Thus, assuming the average Galactic extinction curve ($R_V = 3.1$), the extinction measure is $A_B = 0.069$, $A_r = 0.041$ and $A_K = 0.006 \text{ mag}$. Bottom, the inferred light curve of the transient. The HST datapoint is a direct measure of the OT magnitude. For the other data, the contribution of the host flux has been subtracted for the Gunn-r and K-band fluxes at each epoch. The adopted host galaxy fluxes are $K = 22.1 \pm 0.3 \text{ mag}$ corresponding to $0.9 \pm 0.3 \mu\text{Jy}$ and $0.5 \pm 0.1 \mu\text{Jy}$ (Gunn-r); see below for details. The error bars include photometric error as well as the uncertainties in the estimated host galaxy flux. r-band curve. A single power-law decay model is inconsistent with the r-band data ($\chi^2_{\text{min}}/\text{d.o.f.} \approx 10$, where d.o.f. is degrees of freedom). We then fit the r-band points to a broken power-law model: $F_r = F_*(t/t_{\text{break}})^{\alpha_{1r}}$ for $t \leq t_{\text{break}}$ and $F_r = F_*(t/t_{\text{break}})^{\alpha_{2r}}$ for $t \geq t_{\text{break}}$; here t is the time in days since the GRB and t_{break} is epoch of the break. Using a Levenberg-Marquardt χ^2 minimization⁶², we find the r-band data are adequately fitted by this broken power-law model ($\chi^2 = 12.1$ for 9 d.o.f.) with the following parameters $F_* = 7.0 \pm 1.9 \mu\text{Jy}$, $t_{\text{break}} = 2.04 \pm 0.46 \text{ d}$, $\alpha_{1r} = -1.10 \pm 0.03$, and $\alpha_{2r} = -1.65 \pm 0.06$. The errors quoted are 1σ confidence intervals for each respective parameter. Our value of α_{1r} agrees very well with the determination by Sagar *et al.*⁵³ ($\alpha = -1.10 \pm 0.06$), based on B and R observations obtained over the first two nights. There is some covariance between α_2 and t_{break} so that the confidence level on a subset of parameters may extend beyond the above quoted errors. For instance, concurrent values of ($t_{\text{break}} = 1 \text{ d}$, $\alpha_{2r} = -1.5$) are allowed at the 2σ level, as are ($t_{\text{break}} = 3 \text{ d}$, $\alpha_{2r} = -1.8$). K-band curve. In contrast, the K-band data can be modelled with a single power-law index: $\alpha_K = -1.12 \pm 0.11$ ($\chi^2 = 8.8$ for 6 d.o.f.). This is an acceptable fit (probability of 35%). The K-band data cannot be statistically reconciled to the best-fit r-band model (that is, assuming α_{1r} , α_{2r} and t_{break} to be fixed) with χ^2 ranging from 28 to 66 (7 d.o.f.), depending on the value assumed for the flux of the host galaxy. See Methods for details of flux estimations.

a large value for β is inconsistent with the simple afterglow model. (Supporting this flat spectrum is the observation that the B-band fluxes, although limited, are comparable to the r-band fluxes; see Fig. 2.) Selective extinction by intervening dust in the host galaxy is an unlikely explanation given the strong detection of the afterglow in the B band (Table 1) and in the U band²⁸.

We draw attention to the fact that such a discrepancy between X-ray and optical fluxes is not uncommon and, in our opinion, a similar discrepancy also exists for GRB971214 and GRB980703. This topic is worthy of further study as it is indicative of a fundamental violation of at least one basic assumption of the simple afterglow model (spherical geometry, impulsive energy release and only one dominant emission mechanism—synchrotron emission). Indeed, we show below that the later behaviour of the afterglow is also not consistent with the simplest afterglow model.

Determining the redshift

We obtained spectra of the OT source with the Low Resolution Imaging Spectrometer²⁹ on the Keck II telescope on the night of 23 January 1999. These observations consisted of three spectra, each of about 600-s duration. A quick analysis showed absorption features, indicating that the OT must be at, or beyond, a redshift of 1.61. A subsequent detailed data reduction yielded the spectrum shown in Fig. 3.

The OT spectrum is marked by prominent absorption lines whose suggested identifications are listed in Table 2, and marked in Fig. 3. Similar lines are routinely seen in quasar spectra, and are thought to arise from absorption in intergalactic clouds. Accepting the identi-

fications, we note that all the absorption features can be attributed to a single intergalactic cloud at a redshift $z_{\text{abs}} = 1.6004 \pm 0.0008$. The uncertainty in the redshift has approximately equal contributions from random and systematic errors. The absorption is relatively strong, suggesting that the absorbing system has a high column density of gas. This is only the second time that an absorption redshift has been determined for a GRB OT (the other case being GRB970508², and possibly also GRB980703⁹), thus placing the OT unambiguously at a cosmological distance.

Although the redshift, $z_{\text{abs}} = 1.6$, is a lower limit, it seems likely that the absorption originates in the GRB host galaxy. From the absence of the Lyman- α ($\text{Ly}\alpha$) forest in our spectrum, we can place a firm upper limit to the redshift of $z < 2.9$. The absence of the reported $\text{Ly}\alpha$ forest in the spectrum³⁰ obtained from the Nordic Optical Telescope (NOT) would further lower this limit to $z \lesssim 2.2$. These limits are further supported by the relatively strong continuum detections of the OT in the B band (see Table 1) and in the U band²⁸, which suggest little or no absorption by the intergalactic gas.

No other convincing absorption features (apart from the normal telluric absorption) are detected in our spectrum in the wavelength range $\lambda \approx 4,700\text{--}9,000 \text{ \AA}$. There are no obvious strong emission

Table 1 Photometric observations of GRB990123

Date 1999 (UT)	Telescope	Band	Magnitude	Reference/ observers
Jan. 23.577	P60	r	18.65 ± 0.04	S.C.O.
Jan. 23.578	P200	B	18.93 ± 0.04	L.M.L., J.S.M.
Jan. 23.958	UPSO	R	20.00 ± 0.05	ref. 53
Jan. 24.005	UPSO	B	20.16 ± 0.15	ref. 53
Jan. 24.18	BAO	R	$20.39 \pm 0.15^*$	ref. 54
Jan. 24.194	BAO	B	20.64 ± 0.07	ref. 54
Jan. 24.547	WO	R	$20.93 \pm 0.2^*$	ref. 55
Jan. 24.636	KI	K	18.29 ± 0.04	D. Koerner & D. Kirkpatrick
Jan. 24.934	UPSO	R	$21.22 \pm 0.08^*$	ref. 53
Jan. 24.978	UPSO	B	22.06 ± 0.20	ref. 53
Jan. 25.14	BAO	R	$21.37 \pm 0.15^*$	ref. 54
Jan. 25.940	UPSO	R	$21.73 \pm 0.12^*$	ref. 53
Jan. 26.154	BAO	R	$22.09 \pm 0.10^*$	ref. 54
Jan. 27.652	KI	K	19.73 ± 0.07	M.A.M., I.S.M., H.I.T.
Jan. 27.677	KI	K	20.40 ± 0.08	G. Neugebauer & L.A.
Jan. 30.52	MDM	R	$23.28 \pm 0.18^*$	J.H., I.-A.Y.
Feb. 3.54	MDM	R	$23.88 \pm 0.24^*$	J.H., I.-A.Y.
Feb. 6.6	KI	K	20.84 ± 0.13	N.K.
Feb. 7.610	KI	K	21.00 ± 0.09	S.R.K.
Feb. 8.6	KI	K	20.83 ± 0.11	S.R.K.
Feb. 9.052	HST	R	$24.12 \pm 0.10^\dagger$	ref. 18
Feb. 9.6	KI	K	20.97 ± 0.10	L. Hillenbrand
Feb. 10.6	KI	K	21.21 ± 0.11	L. Hillenbrand
Feb. 9.654	KII	R	$23.91 \pm 0.07^*$	A. Filippenko
Feb. 14.50	MDM	R	$24.10 \pm 0.10^*$	J.H., I.-A.Y.

Telescope acronyms. P60, Palomar 60-inch telescope. P200, Palomar 200-inch Hale telescope. UPSO, UP State Observatory 104-cm telescope. BAO, Bologna Astronomical Observatory 1.5-m telescope. WO, Fred L. Whipple Observatory 1.2-m telescope. KI, Keck-I 10-m telescope. MDM, 2.4-m Hiltner Telescope. KII, Keck-II 10-m telescope. HST, Hubble Space Telescope. Photometric zero-point. The night of 23 January UT was photometric at Palomar. Absolute zero-points for the Gunn-r and B-band data were obtained using the observations of standard star-fields^{56–58}. The absolute zero-point for the K-band observations was obtained using Keck-I data from the night of 24 January UT. Each bandpass zero-point was confirmed independently by at least two members of our group. All subsequent imaging by our group were reduced in the standard manner, and photometry was propagated using a set of secondary stars in the GRB field. Photometry derived from data taken in the Cousins R-band were recalibrated to Gunn-r; the quoted uncertainties include the estimated zero-point and statistical uncertainties. In this table, we include all of our photometric data obtained to date, and also data drawn from the literature which we could reliably place in our B, Gunn-r, and K-band system.

* These are R_c magnitudes corrected to Gunn r-band using secondary reference stars reported in refs 36, 59.

† Observations were conducted with the Space Telescope Imaging Spectrograph (STIS) and CLEAR filter. See ref. 18 for details of conversion of STIS/CLEAR magnitudes to r-band.

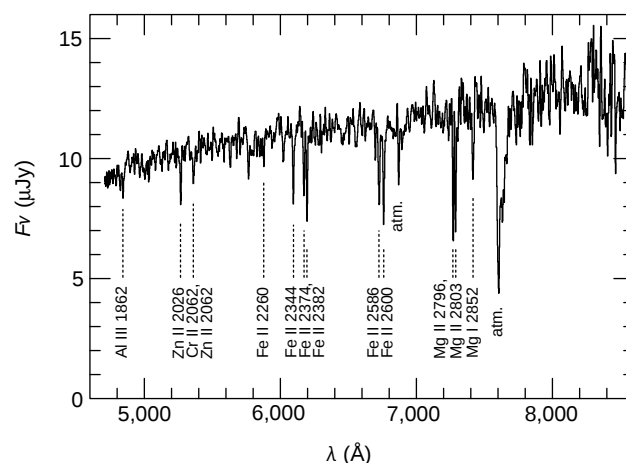


Figure 3 The spectrum of the optical transient of GRB990123. Prominent absorption features are marked; all of these can be attributed to various absorption lines at a redshift, $z_{\text{abs}} = 1.6004$; the two features labelled 'atm.' are telluric absorption features (A and B bands) of the Earth's atmosphere. Three 600-s integrations were taken, though the last one was terminated after 563 s due to the onset of fog; the integrations commenced at UT 15:38, UT 15:49 and UT 16:01, respectively. The last spectrum also suffered from a higher sky background from the brightening dawn sky. The sky conditions appeared to be good during the first two integrations, with seeing ~ 1 arcsec. We used the 300 line mm^{-1} grating and the resulting useful wavelength coverage was 4,700–9,000 Å with an effective resolution of $\sim 11.6 \text{ \AA}$, (full-width at half-maximum). Spectra were taken with the 1.0-arcsec-wide long slit orientated at position angle (PA) 90° (that is, east–west) and centred on the OT; the slit also included the prominent galaxy that is ~ 10 arcsec west of the OT (see Fig. 1). The usual calibration observations were not obtained given the urgency of the observations. The calibration of the wavelength was obtained using a dispersion curve measured with the same grating and slit, but a somewhat different tilt, 10 days earlier. Wavelengths of the unblended night-sky emission lines were then used to adjust the zero-point of the wavelength solution fit; this procedure also compensated flexure of the instrument between the exposures. An approximate flux calibration was accomplished by using the instrument response curve also measured 10 days earlier; the instrument response is stable enough for this approach. The slit losses due to the seeing and the difference of the actual slit PA from the optimal (parallactic) angle at the time of observations make the net flux calibration and the overall the shape of the spectrum somewhat uncertain. Nonetheless, we note that our spectroscopic flux measurement is in excellent agreement with photometry from interpolation of the r-band light curve. In any case, this is not important for the redshift measurements discussed in the text.

lines out to $\lambda \approx 9,600 \text{ \AA}$. The absence of other absorption systems in the OT spectrum is not surprising: in this redshift range there is about one metallic line absorber per unit redshift interval³¹.

The spectrograph slit in these observations also included the bright galaxy ~ 10 arcsec to the west of the OT. We determine a redshift of $z = 0.2783 \pm 0.0005$ for this galaxy, in agreement with the NOT measurement³². From our infrared observations we note that the K-band magnitude of this galaxy is $K = 16.39 \pm 0.03$. At this redshift, this corresponds to a normal, $L \approx L_*$ galaxy; where L_* refers to an average galaxy luminosity in a flux-limited sample.

The energetics of the burst

The combination of a high fluence and a redshift $z \geq 1.6$ imply that this burst is extremely energetic. For this discussion, we assume a standard Friedmann model cosmology with Hubble constant $H_0 = 65 \text{ km s}^{-1} \text{ Mpc}^{-1}$, cosmological density parameter $\Omega_0 = 0.2$, and cosmological constant $\Lambda_0 = 0$ (if $\Lambda_0 > 0$ then the inferred distance would increase, further aggravating the energetics of the burst). At the redshift $z = 1.6004$, the derived luminosity distance is $D_L = 3.7 \times 10^{28} \text{ cm}$, corresponding to the distance modulus $(m - M) = 45.39$, and 1 arcsec in projection corresponds to 8.6 proper kpc, or 22.5 co-moving kpc.

From the observed fluence¹⁴ (energy $> 20 \text{ keV}$) of $5.1 \times 10^{-4} \text{ erg cm}^{-2}$ and the inferred D_L , we estimate the γ -ray energy release, assuming the emission was isotropic, to be $3.4 \times 10^{54} \text{ erg} \approx 1.9 M_\odot c^2$, more than the rest mass energy of a neutron star. The prompt optical flash detected by ROTSE³³ had a V-band magnitude of 8.9, implying a peak luminosity in the ultraviolet band (rest-frame) of $\sim 3.3 \times 10^{16} L_\odot$ (where $L_\odot$ is the solar luminosity). This is about a million times the luminosity of a normal galaxy, and about a thousand times the luminosity of the brightest quasars known. The prompt optical emission is attributed to emission from the reverse shock^{20,34}; in contrast, the late-time afterglow (which is the focus here) arises from the forward shock.

The above energetics and the peak luminosities are truly staggering. Popular models suggest that GRBs are associated with stellar deaths, and not with quasars (or the nuclei of galaxies). Perhaps the strongest observational evidence in favour of this presumption is that some GRBs are found offset from their host galaxy, and are therefore not positionally consistent with an active nucleus. An isotropic energy release of $1.9 M_\odot c^2$ is, however, essentially incompatible with the popular stellar death models (coalescence of neutron stars and the currently popular model of the death of massive stars). This large energy release is barely consistent with exotic models such as that of baryon decay³⁵; in such models essentially the entire rest mass energy of the neutron star is released.

There are two ways to reduce the estimated energy release. The first possibility is that the burst is amplified by gravitational lensing.

Indeed, on the basis of early reports³⁶ of a possible foreground galaxy near the line of sight to the OT, we proposed that GRB990123 was lensed. This hypothesis received support when the NOT team reported³⁰ possible intervening systems at redshifts of 0.286 and 0.201 which could be potential candidates for lensing. Our own subsequent, deeper imaging observations have failed to confirm any foreground galaxy and, as discussed above, there is little spectroscopic evidence for the absorption systems with redshifts of 0.286 and 0.201. There is therefore no direct evidence for lensing at present.

Furthermore, the lensing probability is $\propto A^{-2}$ where A is the amplification provided by the lens. Even with $A \approx 2$, the expected probability for lensing at a redshift of 1.6 is 10^{-3} (ref. 37). This probability is consistent with the observation that roughly 1 in 500 cosmic radio sources, which probably have a similar redshift distribution to GRBs, are lensed. Thus it is not very likely³⁸ that one out of the 15 bursts observed by BeppoSAX so far would be highly magnified.

We now consider the second possibility: the emitting surface in GRB990123 is not spherical. Indeed, almost all energetic sources in astrophysics (such as quasars and accreting stellar black holes, pulsars) are not spherically emitting sources but display jet-like geometry. Not surprisingly there is an extensive literature on jets in the GRB context (see, for example, refs 39–43) as well as for the afterglow emission⁴⁴.

Should the emitting surface be a jet with an opening angle of radius θ_0 then the inferred energy is reduced from the isotropic value by the beaming factor, $f_b \approx \theta_0^2/2$. Here we make the reasonable assumption based on other astrophysical sources that the jet is two-sided. Thus if $\theta_0 \approx 0.3$ then $f_b \approx 5 \times 10^{-2}$, and the γ -ray energy released is $\sim 10^{53} \text{ erg}$, a value within reach of current models for the origin of GRBs.

As discussed below, a marked steepening of the afterglow emission is the clearest and simplest signature of a jet-like geometry. The three well studied GRBs (970228, 970508, 980703)^{26,45–49} exhibit, at late times (epochs longer than few days) a single power-law decay, with indices that are reasonable for spherical expansion, $\alpha \approx -1.2$. (The complicated early time variations seen in afterglows of many GRBs are not relevant to the discussion here.) The radio afterglow of GRB970508 has been used to argue⁵⁰ for beaming in this source, but this result is model-dependent. It is against this backdrop that we now consider if there is evidence for non-spherical geometry in GRB990123.

The shape of the burst: beaming?

Our knowledge of the shape of the emitting region in GRBs is limited because, due to relativistic beaming, only a small portion (angular size $\sim \Gamma^{-1}$, where Γ is the Lorentz factor of the bulk motion of the emitting material) is visible to the observer. Thus the observer is unable to distinguish a sphere from a jet as long as $\Gamma > \theta_0^{-1}$. However, as the source continues its radial expansion, Γ will decrease, and when $\Gamma > \theta_0^{-1}$ there will be a marked decrease in the observed flux. Even if the jet continues to evolve as a cone with a constant opening angle, the observer will see the flux reduced by the ratio of the solid angle of the emitting surface ($\propto \theta_0^2$) to that expected for a spherical expansion ($\propto \Gamma^{-2}$). Thus the light curve will steepen by $\Gamma^2 \theta_0^2 \propto t^{-3/4}$. This is a purely geometric effect.

Rhoads⁵¹ makes the important point that unless the jet is confined by some mechanism (as advocated in ref. 43), it will also expand sideways at the sound speed of shocked relativistic material, $c_s = c/\sqrt{3}$. Thus as the jet evolves, the solid angle of the emitting region increases as $\pi(\theta_0 + c_s t_{\text{co}}/ct)^2$, where t is the time since the burst in the GRB frame, and $t_{\text{co}} = t/\Gamma$ is the elapsed time in the frame moving with the jet. When Γ falls below θ_0^{-1} , the jet spreads in the lateral direction. Consequently, in this regime, Γ decreases exponentially with radius r instead of as a power law. We can therefore take r to be essentially constant during this spreading

Table 2 Absorption lines detected in the spectrum of the OT

Line	$\lambda_{\text{obs,air}}$ (Å)	$\lambda_{\text{rest,vac}}$ (Å)	z	$W_{\lambda_{\text{obs}}}$ (Å)
Al III 1,862	4,843.74	1,862.78	1.6010	1.23 ± 0.07
Zn II 2,026	5,267.29	2,026.14	1.6004	2.08 ± 0.17
Cr II 2,062	5,361.77	2,062.23	1.6607	1.26 ± 0.10
Zn II 2,062	5,361.77	2,062.66	1.6002	1.26 ± 0.10
Fe II 2,260	5,877.17	2,260.78	1.6003	0.91 ± 0.07
Fe II 2,344	6,096.14	2,344.21	1.6012	3.04 ± 0.28
Fe II 2,374	6,173.87	2,373.73	1.6016	3.07 ± 0.28
Fe II 2,382	6,195.29	2,382.76	1.6008	3.60 ± 0.40
Fe II 2,586	6,725.75	2,586.64	1.6009	2.84 ± 0.37
Fe II 2,600	6,759.94	2,600.18	1.6005	3.48 ± 0.14
Mg II 2,796	7,269.47	2,796.35	1.6003	4.59 ± 0.30
Mg II 2,803	7,289.49	2,803.53	1.6008	4.80 ± 0.52
Mg I 2,852	7,416.97	2,852.97	1.6005	2.87 ± 0.18

$\lambda_{\text{obs,air}}$ is the wavelength of a line as measured in the spectrograph. $\lambda_{\text{rest,vac}}$ is the wavelength of the same line as measured in the frame of the source, and in vacuum. $W_{\lambda_{\text{obs}}}$ is the observed line equivalent width; for the rest-frame values, divide by $(1+z)$. The observed absorption line at 5,362 Å is a blend of two lines, Cr II 2,062 and Zn II 2,062, and thus it appears twice.

phase. Since $t_E \approx r/2\Gamma^2 c$, we find $\Gamma \propto t_E^{-1/2}$ during the spreading phase; here, t_E is the elapsed time in the frame of the observer at Earth. With this scaling the typical synchrotron frequency, ν_m , decreases as t_E^{-2} and the flux at this frequency drops as t_E^{-1} . The flux at a fixed frequency above ν_m drops as $F_\nu = F_{\nu_m}(\nu/\nu_m)^{-(p-1)/2} \propto t_E^{-p}$, where $p \approx 2.5$ is the electron energy power-law index. This is more than one power of t steeper than that for a spherical afterglow, $t_E^{-3(p-1)/4} \propto t_E^{-1.1}$.

In contrast to the above discussion, we expect another kind of break, and that is when the electrons responsible for the photons in the band of interest enter the cooling regime. In this case, α decreases by 1/4 (see, for example, ref. 23). However, unlike the break due to jet geometry (discussed above) this break is not broadband. But regardless of whether electron cooling or jets are involved, α is expected to evolve over a timescale of t_{break} . Thus the expected change in α , $\Delta\alpha \equiv \alpha_1 - \alpha_2$ is 3/4 (constant θ_0), $1 - \alpha_1/3$ (spreading jet) or 1/4 (electron cooling), should be considered to be upper limits; see also ref. 52.

As can be seen from Fig. 2 we do indeed see a break in the r-band flux with $\Delta\alpha_r = 0.55 \pm 0.07$; the 95% confidence interval, taking account of covariance with the epoch of the break, is 0.4–0.7 (see Fig. 2). On statistical grounds, we can rule out $\Delta\alpha = 1/4$. The prediction of $\Delta\alpha = 1/4$ for electron cooling arises from basic synchrotron theory and is therefore a robust prediction. Therefore, we firmly conclude that the observed break is not due to electron cooling.

The measured value of $\Delta\alpha$ supports the hypothesis of a jet in GRB990123. However, the lack of a break in the K-band light curve is an issue of considerable concern. As explained in Fig. 2 legend, a single power-law model provides a good fit to the K-band data. Within statistical errors, we can state that the r-band model does not conform to the K-band light curve. Unfortunately, the quality and quantity of the K-band data and the uncertainty in the adopted flux of the host galaxy do not permit us to derive parameters for a broken power law model independent of those derived from the r-band light curve. (Future K-band observations when the OT has essentially disappeared would help us resolve at least the latter uncertainty.)

We cautiously advance the hypothesis that we are seeing evidence for a jet (perhaps with no sideways expansion). If so, $\theta_0 \approx \Gamma(t_{\text{break}})^{-1}$ and for typical parameters, $\theta_0 \approx 0.2$, in which case $f_b \approx 0.02$. Thus the energy released in γ -rays alone is 6×10^{52} erg. Including the X-ray afterglow¹³ raises the total energy release to 10^{53} erg. GRB990123 was indeed a very energetic GRB. Following the discovery of the afterglow phenomenon^{5–7}, it was thought that afterglow observations would provide only global parameters of the GRB explosion (such as energy released, and ambient density), but it now appears that afterglow observations may also give us insight into the geometry of the explosion. $\square$

Methods

The flux of the host galaxy was estimated as follows. For the r-band observations, we used the HST/STIS imaging where the host galaxy is clearly resolved and thus the flux of the OT and the host can be measured quite accurately. In ref. 18 we discuss in detail the tie between Gunn-r photometry and the STIS photometry (done with the CLEAR filter which corresponds approximately to the V band) as well as R-band photometry. The photometric ties are robust, so that the flux of the host in r band is well determined, $0.5 \pm 0.1 \mu\text{Jy}$.

We determined the flux of the host galaxy in K-band as follows. We used the deepest exposures with the best seeing images (9 and 10 February). In these images the OT is clearly resolved from the host galaxy. Sets of pixels dominated by the OT or by the galaxy were masked, and total fluxes with such censored data are evaluated in photometric apertures of varying radii. Total fluxes of the OT + galaxy were also measured in the same apertures using the uncensored data. We also varied the aperture radii, and the position and the size of the sky measurement annulus. On February 9 (February 10) UT we find the OT

contributes 65 (57) per cent (± 10 per cent) of the total OT + galaxy light. The estimated errors of the fractional contributions of the OT to the total light reflect the scatter obtained from variations in the parameters of these image decompositions. In both epochs the fractional contribution of the host implies a flux of the host galaxy is $0.9 \pm 0.3 \mu\text{Jy}$ ($K_{\text{host}} = 22.1 \pm 0.3$ mag). As a further test, we fitted the K-band light curve assuming the flux is given as the sum of a power-law (OT) plus a constant flux (host galaxy). This fit yields $K_{\text{host}} = 21.55 \pm 0.20$ mag, in agreement with the magnitude derived from derived direct imaging. However, we note that the fit is barely acceptable ($\chi^2 = 11.9$, 5 d.o.f.).

Received 12 February; accepted 1 March 1999.

- Klebesadel, R. W., Strong, I. B. & Olson, R. A. Observations of gamma-ray bursts of cosmic origin. *Astrophys. J.* **182**, L85–L88 (1973).
- Metzger, M. R. et al. Spectral constraints on the redshift of the optical counterpart to the γ -ray burst of May 8, 1997. *Nature* **387**, 878–879 (1997).
- Boella, G. et al. BeppoSAX, the wide band mission for x-ray astronomy. *Astron. Astrophys. Suppl. Ser.* **122**, 299–399 (1997).
- Levine, A. M. et al. First results from the All-Sky Monitor on the Rossi X-ray Timing Explorer. *Astrophys. J.* **469**, L33–L36 (1996).
- Costa, E. et al. Discovery of an X-ray afterglow associated with the γ -ray burst of 28 February 1997. *Nature* **387**, 783–785 (1997).
- van Paradijs, J. et al. Transient optical emission from the error box of the γ -ray burst of 28 February 1997. *Nature* **386**, 686–689 (1997).
- Frail, D. A., Kulkarni, S. R., Nicastro, L., Feroci, M. & Taylor, G. B. The radio afterglow from the γ -ray burst of 8 May 1997. *Nature* **389**, 261–263 (1997).
- Kulkarni, S. R. et al. Identification of a host galaxy at redshift $z = 3.42$ for the γ -ray burst of 14 December 1997. *Nature* **393**, 35–39 (1998).
- Djorgovski, S. G. et al. Spectroscopy of the host galaxy of the gamma-ray burst 980703. *Astrophys. J.* **508**, L17–L20 (1998).
- Djorgovski, S. G. et al. GRB980613: Spectroscopy of the host galaxy. *GCN Circ.* No. 189 (1999).
- Grindlay, J. Fast X-ray transients and gamma-ray bursts: constraints on beaming. *Astrophys. J.* **510**, 710–714 (1999).
- Greiner, J., Voges, W., Boiler, T. & Hartmann, D. Search for GRB afterglows in the ROSAT all-sky survey. *Astron. Astrophys.* (submitted).
- Heise, J. et al. A very energetic γ -ray burst on January 23 1999 and its X-ray afterglow. *Nature* (submitted).
- Kippen, R. M. GRB990123: BATSE observations. *GCN Circ.* No. 224 (1999).
- Piro, L. GRB990123, BeppoSAX WFC detection and NFI planned follow-up. *GCN Circ.* No. 199 (1999).
- The GRB Coordinates Network at (<http://gcn.gsfc.nasa.gov/gcn/>).
- Beckwith, S. For immediate posting to GCN—HST Data GRB990123 available. *GCN Note* No. 254 (1999).
- Bloom, J. S. et al. The host galaxy of GRB990123. Preprint astro-ph/9902182 at (<http://xxx.lanl.gov>) (1999).
- Katz, J. I. Low-frequency spectra of gamma-ray bursts. *Astrophys. J.* **432**, L110–L113 (1994).
- Mészáros, P. & Rees, M. J. Optical and long-wavelength afterglow from gamma-ray bursts. *Astrophys. J.* **476**, 232–240 (1997).
- Vietri, M. The soft X-ray afterglow of gamma-ray bursts, a stringent test for the fireball model. *Astrophys. J.* **478**, L9–L12 (1997).
- Waxman, E. Gamma-ray-burst afterglow: supporting the cosmological fireball model, constraining parameters, and making predictions. *Astrophys. J.* **485**, L5–L8 (1997).
- Sari, R., Piran, T. & Narayan, R. Spectra and light curves of gamma-ray burst afterglows. *Astrophys. J.* **497**, L17–L20 (1998).
- Wijers, R. A. M. J. & Galama, T. J. Physical parameters of GRB970508 and GRB971214 from their afterglow synchrotron emission. Preprint astro-ph/9806175 at (<http://xxx.lanl.gov>) (1998).
- Ramaprakash, A. N. et al. The energetic afterglow of the γ -ray burst of 14 December 1997. *Nature* **393**, 43–46 (1998).
- Sokolov, V. V. et al. BVR₁₂ photometry of GRB970508 optical remnant: May–August 1997. *Astron. Astrophys.* **334**, 117–123 (1998).
- Groot, P. et al. The decay of optical emission from the γ -ray burst GRB970228. *Nature* **387**, 479–481 (1997).
- Falco, E., Petry, C., Impey, C., Koekemoer, A. & Rhoads, J. GRB990123 optical observations. *GCN Circ.* No. 214 (1999).
- Oke, J. B. et al. The Keck Low-Resolution Imaging Spectrometer. *Publ. Astron. Soc. Pacif.* **107**, 375–385 (1995).
- Hjorth, J. et al. GRB990123 spectroscopic redshifts. *GCN Circ.* No. 219 (1999).
- Steidel, C. C. & Sargent, W. L. W. MgII absorption in the spectra of 103 QSOs: implications for the evolution of gas in high-redshift galaxies. *Astrophys. J. Suppl. Ser.* **80**, 1–108 (1992).
- Hjorth, J. et al. GRB990123 NOT spectrum update. *GCN Circ.* No. 249 (1999).
- Akerlof, C. et al. Observation of contemporaneous optical radiation from a γ -ray burst. *Nature* **398**, 400–402 (1999).
- Sari, R. & Piran, T. GRB990123, The optical flash and the fireball model. Preprint astro-ph/9902009 at (<http://xxx.lanl.gov>) (1999).
- Pen, U.-L. & Loeb, A. Gamma-ray bursts from baryon decay in neutron stars. *Astrophys. J.* **509**, 537–543 (1998).
- Djorgovski, S. G. et al. *GCN Circ.* No. 216 (1999).
- Holz, D. E., Coleman, M. M. & Quashnock, J. M. Gravitational lensing limits on the average redshift of gamma-ray bursts. *Astrophys. J.* **510**, 54–63 (1999).
- Schaefer, B. E. GRB990123, probability of gravitational lensing. *GCN Circ.* No. 241 (1999).
- Mochkovich, R., Hernanz, M., Isern, J. & Martin, X. Gamma-ray bursts as collimated jets from neutron star/black hole mergers. *Nature* **361**, 236–237 (1993).
- Mao, S. & Yi, I. Relativistic beaming and gamma-ray bursts. *Astrophys. J.* **424**, L131–L134 (1994).
- Katz, J. I. Yet another model for gamma ray bursts. *Astrophys. J.* **490**, 633–640 (1997).
- Mészáros, P. & Rees, M. J. Poynting jets from black holes and cosmological gamma-ray bursts. *Astrophys. J.* **482**, L29–L32 (1997).
- Dar, A. Can fireball models explain gamma-ray bursts? *Astrophys. J.* **500**, L93–L96 (1998).
- Rhoads, J. E. How to tell a jet from a balloon: a proposed test for beaming in gamma-ray bursts. *Astrophys. J.* **478**, L1–L4 (1997).
- Pedersen, H. et al. Evidence for diverse optical emission from gamma-ray burst sources. *Astrophys. J.* **496**, 311–315 (1998).

46. Galama, T. *et al.* Optical follow up of GRB970508. *Astrophys. J.* **497**, L13–L16 (1998).
47. Bloom, J. S., Djorgovski, S. G., Kulkarni, S. R. & Frail, D. A. The host galaxy of GRB970508. *Astrophys. J.* **507**, L25–L28 (1998).
48. Fruchter, A. S. *et al.* The fading optical counterpart of GRB970228, six months and one year later. Preprint astro-ph/9807295 at (<http://xxx.lanl.gov>) (1998).
49. Frail, D. A., Bloom, J. S., Kulkarni, S. R. & Taylor, G. B. The light curve and the spectrum of the radio afterglow of GRB980703. *Astrophys. J.* (submitted).
50. Waxman, E., Kulkarni, S. R. & Frail, D. A. Implications of the radio afterglow from the gamma-ray burst of 1997 May 8. *Astrophys. J.* **497**, 288–293 (1998).
51. Rhoads, J. E. Constraining gamma ray burst beaming. Preprint, Kitt Peak National Observatory (1998).
52. Mészáros, P. & Rees, M. J. GRB990123: reverse and internal shock flashes and late afterglow behavior. Preprint astro-ph/9902367 at (<http://xxx.lanl.gov>) (1999).
53. Sagar, R. *et al.* Optical follow up of the GRB990123 source from UPSO, Nainital. Preprint astro-ph/9902196 at (<http://xxx.lanl.gov>) (1999).
54. Masetti, N. *et al.* GRB990123, optical BVRI observations. *GCN Circ.* No. 233 (1999).
55. Garnavich, P., Jha, S., Stanek, K. & Garcia, M. GRB990123, optical observation. *GCN Circ.* No. 215 (1999).
56. Thuan, T. X. & Gunn, J. E. A new four-colour intermediate-band photometric system. *Proc. Astron. Soc. Pacif.* **88**, 543–547 (1976).
57. Oke, J. B. & Gunn, J. E. Secondary standard stars for absolute spectrophotometry. *Astrophys. J.* **266**, 713–717 (1983).

58. Landolt, A. UBVR photometric standard stars in the magnitude range 11.5–16.0 around the celestial equator. *Astron. J.* **104**, 340–376 (1992).
59. Zhu, J. & Zhang, H. T. GRB990123 optical observation. *GCN Circ.* No. 204 (1999).
60. Monet, D. G. The 526,280,881 objects in the USNO-A2.0 catalog. *Bull. Am. Astron. Soc.* **193**, 225 (1998).
61. Schlegel, D. J., Finkbeiner, D. P. & Davis, M. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *Astrophys. J.* **500**, 525–553 (1998).
62. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes in C* 683–699 (Cambridge Univ. Press, New York, 1992).

Acknowledgements. We thank G. Neugebauer for obtaining infrared data on 29 January 1999. We thank A. Filippenko, L. Hillenbrand, J. Carpenter and K. Leighly for agreeing to exchange telescope time, thereby enabling us to follow up this GRB, and B. Oppenheimer for help with observations. Some of the observations reported here were obtained at the W. M. Keck Observatory, which is operated by the California Association for Research in Astronomy, a scientific partnership among California Institute of Technology, the University of California and NASA. It was made possible by the support of the W. M. Keck Foundation. S.R.K.'s research is supported by the NSF and NASA. S.G.D. acknowledges partial support from the Bressler Foundation.

Correspondence and requests for materials should be addressed to S.R.K. (e-mail: srk@astro.caltech.edu).

The effect of magnetic fields on γ -ray bursts inferred from multi-wavelength observations of the burst of 23 January 1999

T. J. Galama¹, M. S. Briggs², R. A. M. J. Wijers³, P. M. Vreeswijk¹, E. Rol¹, D. Band⁴, J. van Paradijs^{1,2}, C. Kouveliotou^{5,6}, R. D. Preece², M. Bremer⁷, I. A. Smith⁸, R. P. J. Tilanus⁹, A. G. de Bruyn^{10,11}, R. G. Strom^{1,10}, G. Pooley¹², A. J. Castro-Tirado^{13,14}, N. Tanvir^{15,16}, C. Robinson¹⁷, K. Hurley¹⁸, J. Heise¹⁹, J. Telting²⁰, R. G. M. Rutten²⁰, C. Packham²⁰, R. Swaters¹¹, J. K. Davies⁹, A. Fassia²¹, S. F. Green²², M. J. Foster²², R. Sagar²³, A. K. Pandey²³, Nilakshi²³, R. K. S. Yadav²³, E. O. Ofek²⁴, E. Leibowitz²⁴, P. Ibbetson²⁴, J. Rhoads²⁵, E. Falco²⁶, C. Petry²⁷, C. Impey²⁷, T. R. Geballe²⁸ & D. Bhattacharya²⁹

Gamma-ray bursts (GRBs) are thought to arise when an extremely relativistic outflow of particles from a massive explosion (the nature of which is still unclear) interacts with material surrounding the site of the explosion. Observations of the evolving changes in emission at many wavelengths allow us to investigate the origin of the photons, and so potentially determine the nature of the explosion. Here we report the results of γ -ray, optical, infrared, submillimetre, millimetre and radio observations of the burst GRB990123 and its afterglow. Our interpretation of the data indicates that the initial and afterglow emissions are associated with three distinct regions in the fireball. The peak flux of the afterglow, one day after the burst, has a lower frequency than observed for other bursts; this explains the short-lived radio emission. We suggest that the differences between bursts reflect variations in the magnetic-field strength in the afterglow-emitting regions.

The current 'standard' model for GRBs and their afterglows is the fireball-plus-blastwave model (see ref. 1 for a review). It invokes the release of a large amount of energy, of the order of $M_{\odot}c^2$, into a volume less than 1 light-ms across (here $M_{\odot}$ is the solar mass and c is the velocity of light). This leads to a 'fireball' that expands ultra-relativistically (Lorentz factor $\Gamma \geq 300$); when it runs into the surrounding medium a 'forward shock' ploughs into the medium and heats it, and a 'reverse shock' does the same to the ejecta. The

GRB itself is thought to owe its multi-peaked light curve to a series of 'internal shocks' that develop in the relativistic ejecta before they collide with the ambient medium. The heated gas is presumed to form relativistic electrons and magnetic fields with energy densities near equipartition, that is, similar to that of the gas, and then emit synchrotron radiation². As the forward shock is weighed down by increasing amounts of swept-up material it becomes less relativistic, and produces a slowly fading 'afterglow' of X rays, then ultraviolet/

¹ Astronomical Institute 'Anton Pannekoek', University of Amsterdam, & Center for High Energy Astrophysics, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands; ² Physics Department, University of Alabama in Huntsville, Huntsville, Alabama 35899, USA; ³ Department of Physics and Astronomy, SUNY Stony Brook, New York 11794-3800, USA; ⁴ CASS, University of California in San Diego, La Jolla, California 92093, USA; ⁵ Universities Space Research Association; ⁶ NASA/MSFC, Code ES-84, Huntsville, Alabama 35812, USA; ⁷ Institut de Radio Astronomie Millimétrique, 300 rue de la Piscine, F-38400 Saint-Martin d'Hères, France; ⁸ Department of Space Physics and Astronomy, Rice University, MS-108, 6100 South Main, Houston, Texas 77005-1892, USA; ⁹ Joint Astronomy Centre, 660 North A'ohoku Place, University Park, Hilo, Hawaii 96720, USA; ¹⁰ NFRA, Postbus 2, 7990 AA Dwingeloo, The Netherlands; ¹¹ Kapteyn Astronomical Institute, Postbus 800, 9700 AV, Groningen, The Netherlands; ¹² Mullard Radio Astronomy Observatory, Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge CB3 0HE, UK; ¹³ Laboratorio de Astrofísica Espacial y Física Fundamental (LAEFF-INTA), PO Box 50727, E-28080 Madrid, Spain; ¹⁴ Instituto de Astrofísica de Andalucía (IAA-CSIC), PO Box 03004, E-18080 Granada, Spain; ¹⁵ Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK; ¹⁶ Department of

Physical Sciences, University of Hertfordshire, College Lane, Hatfield, Herts AL10 9AB, UK; ¹⁷ National Science Foundation; ¹⁸ University of California at Berkeley, Space Sciences Laboratory, Berkeley, California 94720-7450, USA; ¹⁹ Space Research Organisation Netherlands (SRON), Sorbonnelaan 2, 3584 CA Utrecht, The Netherlands; ²⁰ Isaac Newton Group, Apartado de Correos, 321, 38780 Santa Cruz de La Palma, Islas Canarias, Spain; ²¹ Astrophysics Group, Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ, UK; ²² Unit for Space Sciences and Astrophysics, School of Physical Sciences, Physics Laboratory, University of Kent at Canterbury, Canterbury, Kent CT2 7NR, UK; ²³ U.P. State Observatory, Manora Peak, Nainital-263129, India; ²⁴ Wise Observatory, Tel Aviv University, Ramat Aviv, Tel Aviv 699 78, Israel; ²⁵ Kitt Peak National Observatory, 950 N. Cherry Avenue, PO Box 26732, Tucson, Arizona 85726, USA; ²⁶ Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA; ²⁷ Steward Observatory, University of Arizona, Tucson Arizona 85719, USA; ²⁸ Gemini Observatory, 670 N. A'ohoku Place, University Park, Hilo, Hawaii 96720, USA; ²⁹ Raman Research Institute, Bangalore 560 080, India.

optical/infrared radiation, and then millimetre-wave and radio radiation.

Mészáros and Rees³, and Sari and Piran⁴, pointed out that when the external shock first forms, optical emission of about ninth–fourteenth magnitude from a reverse shock will be visible during, or soon after, the high-energy burst produced by the internal shocks. Other models have been put forward which lead to an optical brightness of about eighteenth magnitude during the GRB⁵. Multi-wavelength observations during, and immediately after, the burst are necessary to detect the emissions from the different regions postulated by these models and transitions between them.

The X-ray emission observed with the BeppoSAX satellite during the ‘tails’ of GRBs is consistent with extrapolations backwards in time of the X-ray afterglow detected many hours later^{6–10}, suggesting that the prompt X-ray emission merges smoothly into the afterglow. On the other hand, the HEAO-1 satellite observed a prompt X-ray signal in GRB780506 which disappeared and then reappeared after a few minutes, suggesting a gap between the prompt emission and the afterglow¹¹.

The ground-based ROTSE (Robotic Optical Transient Search Experiment) detection¹² of prompt optical emission during and immediately after GRB990123 shows that optical observations can address the question of which physical region is radiating, and when. The first ROTSE exposure occurred during the main peak of the burst; weaker γ -ray emission was also present during the second and third ROTSE exposures, but was no longer detected in subsequent exposures.

We here present an extensive set of multi-wavelength observations for GRB990123, which cover the γ -ray to radio range and permit an analysis of the relation of the burst’s prompt emission to its afterglow in the context of the ‘fireball’ model.

Observations

Prompt γ -ray emission. In Fig. 1 we show the light curves of GRB990123 in two energy bands, obtained by the Burst and Transient Source Experiment (BATSE) on board the CGRO satellite. The burst profile is dominated by two peaks, each lasting about 8

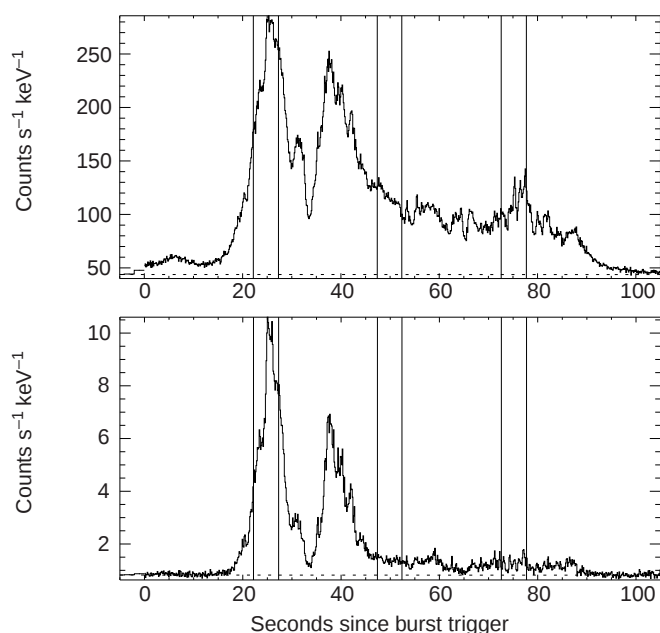


Figure 1 The BATSE light curves of GRB990123 in two energy ranges, 25–230 keV (top) and 320–1,800 keV (bottom). Times are relative to the BATSE trigger time of 09:46:56.1 on January 23. The vertical lines indicate the time intervals of the first three ROTSE observations¹².

seconds (full-width at half-maximum, FWHM), separated by 12 seconds, followed by a shoulder lasting some 40 seconds. The low-energy emission persists longer than the high-energy emission, that is, GRB990123 presents a typical case of ‘hard-to-soft’ spectral evolution¹³. Particularly striking is the paucity of >300 -keV emission during the shoulder. The T_{90} duration of the burst¹⁴, as measured in the 50–300 keV range, is 63.30 ± 0.26 seconds.

Using the BATSE data, we find for the peak flux (50–300 keV) and total fluence (>20 keV) values of $F_{\text{max}} = 3.8 \times 10^{-6} \text{ erg cm}^{-2} \text{ s}^{-1}$ and $E_b = 3.0 \times 10^{-4} \text{ erg cm}^{-2}$, respectively. The redshift ($z = 1.61$) of the optical afterglow, obtained from Fe and Mg absorption lines¹⁵, implies a minimum luminosity distance of 11 Gpc, and a total rest-frame (>20 keV) γ -ray energy of $2 \times 10^{54} \text{ erg}$, using a Hubble constant $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, a cosmological density parameter $\Omega_0 = 0.3$, and assuming isotropic emission. To investigate the relation between the optical emission detected with ROTSE and the GRB we made spectral fits to the γ -ray data during the time intervals corresponding to the first three ROTSE images. In Fig. 2 we compare our γ -ray fits to the simultaneous optical observations by ROTSE. Extending the low-energy power-law to optical frequencies, even allowing for the slope uncertainties, we find that in all three spectra the optical data points fall well above the low-energy extension of the γ -ray spectrum, independent of the slope of the latter.

Multi-wavelength afterglow emission. We observed the afterglow of GRB990123 in optical, infrared, submillimetre, millimetre and radio passbands (Tables 1, 2 and 3). The optical R-band light curve after the first 0.1 day is well described by a power-law decay, $F_\nu = F_0 t^\alpha$ (where t is the time since the burst): we find $\alpha = -1.12 \pm 0.03$ (reduced chi-square, $\chi_r^2 = 30/28$). To account for the presence of light from the underlying host galaxy¹⁶ we have also fitted a model consisting of a power-law decay plus a constant flux; we find $\alpha = -1.14 \pm 0.03$ and the R magnitude of the host galaxy, $R_{\text{host}} > 24.9$ ($\chi_r^2 = 30/27$). The infrared observations show a decay that is consistent with that in the R band; with decay constants in the H and K bands, $\alpha_H = -0.92 \pm 0.22$ (from ref. 17) and $\alpha_K = -1.14 \pm 0.08$, respectively. The late-time power law decay

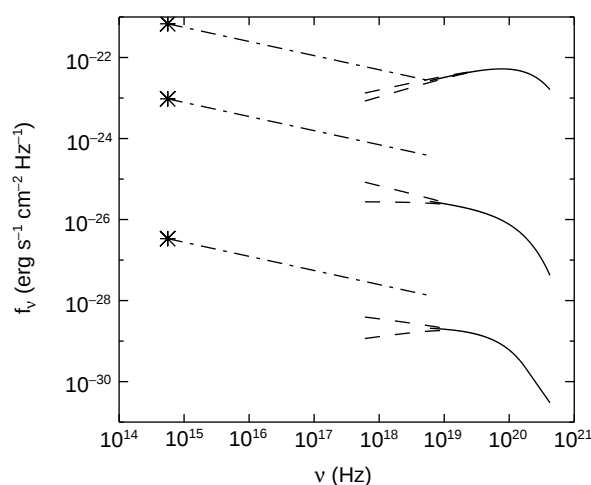


Figure 2 Optical flux and γ -ray spectra for the time intervals of the three ROTSE observations. The γ -ray curves (solid) are fits to BATSE spectra using the Band function³⁶, which consists of a low-energy power law with an exponential cut-off which merges smoothly with a high-energy power law. The curves for the first (third) time interval have been shifted up (down) by a factor of 1,000. The data point on the left is the reported ROTSE magnitude converted into a flux density in the middle of the V band. The dashed curves on the low-energy end of the γ -ray spectra show the 3σ variations in the low-energy spectral index. The dot-dashed curves are the extrapolation of the optical flux to the X-ray band using the power law which connects the observed optical and lowest-energy γ -ray fluxes for the first ROTSE time interval.

connects smoothly with the last three data points of the ROTSE observations (Fig. 3), but not with the first three: the slope of the light curve before these three points is much steeper.

We detect the radio afterglow at 4.88 GHz at the improved optical position¹⁸, using the Westerbork radio telescope. An r.m.s.-noise weighted flux density average gives $F_{4.88\text{GHz}} = 118 \pm 40 \mu\text{Jy}$ (average epoch January 24.46 UT). After January 26 we do not detect the source at 4.88 GHz ($F_{4.88\text{GHz}} = -7 \pm 31 \mu\text{Jy}$; weighted average of the January 26, 28 and 29 data). Similar behaviour was observed¹⁹ at 8.46 GHz: initially the source was not detected ($<68 \mu\text{Jy}$; 2σ ; January 23.63 UT), then it was detected ($260 \pm 32 \mu\text{Jy}$; January 24.65 UT), and then not detected after January 26. Such radio behaviour is unique, both for its early appearance as well as its rapid decline. The source was not detected at 1.38, 15, 85, 140, 225, 350 and 670 GHz (Table 2).

We have reconstructed the radio to X-ray afterglow spectrum on January 24.65 UT (Fig. 4). This interval was selected for the long wavelength coverage possible at the only time of radio detections. Describing the spectrum by a power law ($F_\nu \propto \nu^\beta$) we find that in the optical range $\beta = -0.75 \pm 0.23$, between the optical and X-ray wavebands the spectral slope is $\beta = -0.67 \pm 0.02$, while in the radio range the spectrum rises as $\beta = +1.4 \pm 0.7$.

Discussion

Of the six GRBs with known redshifts, there are two with total energy (assuming isotropic emission) in excess of 10^{54} erg; GRB990123, discussed here, and GRB980329 (ref. 20). Of course, beaming of the γ -ray emission (that is, strongly anisotropic emission) may make the total energy substantially smaller, and until we have a good understanding of beaming the total energy is not a severe constraint on theoretical models of the initial explosion.

If one assumes that the emission detected by ROTSE comes from a non-relativistic source of size ct , then we can obtain a lower limit

Table 1 Optical and infrared observations of GRB990123

UT date (1999)	Magnitude	Filter	Exp. time (s)	Telescope/ reference
Jan. 23.970	19.58 $\pm$ 0.14	R	3,600	UPSO 1-m
Jan. 24.028	20.42 $\pm$ 0.10	B	3,000	UPSO 1-m
Jan. 24.999	21.82 $\pm$ 0.12	B	3,600	UPSO 1-m
Jan. 25.029	21.14 $\pm$ 0.16	V	600	UPSO 1-m
Jan. 24.066	20.38 $\pm$ 0.26	B	120	Wise 1-m
Jan. 24.062	19.94 $\pm$ 0.13	R	600	Wise 1-m
Jan. 24.081	20.08 $\pm$ 0.11	R	600	Wise 1-m
Jan. 24.110	20.14 $\pm$ 0.14	R	600	Wise 1-m
Jan. 24.116	20.02 $\pm$ 0.09	R	600	Wise 1-m
Jan. 25.079	20.82 $\pm$ 0.15	R	900	Wise 1-m
Jan. 25.092	21.75 $\pm$ 0.21	B	1,200	Wise 1-m
Jan. 25.105	20.95 $\pm$ 0.16	R	900	Wise 1-m
Jan. 24.198	20.12 $\pm$ 0.06	R	1,000	JKT
Jan. 24.210	20.14 $\pm$ 0.05	R	1,000	JKT
Jan. 24.222	20.16 $\pm$ 0.06	R	1,000	JKT
Jan. 24.456	20.60 $\pm$ 0.06	R	900	WIYN
Jan. 24.477	20.55 $\pm$ 0.04	R	600	WIYN
Jan. 24.487	20.14 $\pm$ 0.08	I	600	WIYN
Jan. 24.496	20.89 $\pm$ 0.06	V	600	WIYN
Jan. 24.505	21.21 $\pm$ 0.04	B	600	WIYN
Jan. 24.515	20.57 $\pm$ 0.04	R	600	WIYN
Jan. 24.544	20.96 $\pm$ 0.08	V	600	WIYN
Jan. 24.553	20.58 $\pm$ 0.05	R	600	WIYN
Feb. 3.519	23.57 $\pm$ 0.35	R	900	WIYN
Feb. 3.533	>22.8	I	900	WIYN
Feb. 7.513	23.59 $\pm$ 0.15	R	3,500	WIYN
Feb. 8.460	24.00 $\pm$ 0.15	R	3,600	WIYN
Jan. 24.497	20.34 $\pm$ 0.10	U	900	KPNO 4-m
Jan. 24.660	19.14 $\pm$ 0.13	H	1,020	UKIRT
Jan. 27.660	20.39 $\pm$ 0.26	H	1,080	UKIRT

Eight reference stars (see Table 3) were used to obtain the differential magnitude of the optical transient (OT) in each observation. Telescopes: WIYN 3.5-m telescope at Kitt Peak (WIYN), Uttar Pradesh State Observatory 1-m telescope (UPSO 1-m), Wise Observatory 1-m telescope (Wise 1-m), Jakobus Kapteyn Telescope, La Palma (JKT), Kitt Peak National Observatory 4-m Mayall telescope (KPNO 4-m), United Kingdom Infra-Red Telescope, Hawaii (UKIRT). Instruments and CCDs used: UPSO 1-m: CCD camera, 512×512 CCD, $0.76''/\text{pixel}^{-1}$; Wise 1-m: TEK 1k $\times$ 1k CCD, $0.70''/\text{pixel}^{-1}$; JKT: TEK 2k $\times$ 2k CCD, $0.33''/\text{pixel}^{-1}$; WIYN: TEK 2k $\times$ 2k CCD, $0.197''/\text{pixel}^{-1}$; KPNO 4-m: 8k $\times$ 8k CCD mosaic camera, $0.258''/\text{pixel}^{-1}$; UKIRT: IRCAM3 with FPA42 256 $\times$ 256 detector, $0.29''/\text{pixel}^{-1}$.

to the brightness temperature T_b because the rest-frame T_b cannot exceed the Compton limit of 10^{12} K. The observed $T_b \geq 10^{17}$ K is a factor Γ^3 times the rest-frame value, which means that the bulk Lorentz factor, Γ , is ≥ 50 , confirming the highly relativistic nature of the GRB source. In view of this, and the independent evidence^{2,21,22} that GRBs and their afterglows are emitted by relativistic shocks, we will discuss our results in terms of this 'fireball' model.

The ROTSE observations¹² have shown that the prompt optical and γ -ray light curves do not track each other. We show that the prompt optical emission detected with ROTSE during GRB990123

Table 2 Radio and (sub) millimetre observations of GRB990123

Freq (GHz)	UT date 1999 (central)	Duration (h)	Flux (μJy)	Reference
1.380	Jan. 27.20	8.4	37 $\pm$ 22	WSRT
4.88	Jan. 24.28	12	104 $\pm$ 42	WSRT
4.88	Jan. 25.47	2.9	164 $\pm$ 100	WSRT
4.88	Jan. 26.29	1.4	-65 $\pm$ 130	WSRT
4.88	Jan. 28.27	12	-3 $\pm$ 41	WSRT
4.88	Jan. 29.17	7.4	-9.3 $\pm$ 55	WSRT
4.88	Jan. 30.27	12	-74 $\pm$ 44	WSRT
4.88	Jan. 31.26	12	20 $\pm$ 41	WSRT
4.88	Feb. 6.25	12	30 $\pm$ 44	WSRT
4.88	Feb. 7.13	6.7	112 $\pm$ 85	WSRT
8.46	Jan. 23.63		<68 (2σ)	VLA
8.46	Jan. 24.65		260 $\pm$ 32	VLA
8.46	Jan. 26		<78 (2σ)	VLA
8.46	Jan. 27		<50 (2σ)	VLA
8.46	Jan. 28		<50 (2σ)	VLA
15.0	Jan. 25.26	9.9	160 $\pm$ 180	Ryle
15.0	Jan. 26.27	10.4	-12 $\pm$ 180	Ryle
15.0	Jan. 29.20	6.4	-197 $\pm$ 200	Ryle
15.0	Jan. 30.40	3.7	106 $\pm$ 420	Ryle
15.0	Jan. 31.39	3.8	48 $\pm$ 300	Ryle
15.0	Feb. 4.39	3.6	378 $\pm$ 270	Ryle
15.0	Feb. 6.37	4.1	-140 $\pm$ 300	Ryle
15.0	Feb. 8.36	2.8	-16 $\pm$ 280	Ryle
86.27	Jan. 25.18	4.5	(1.0 $\pm$ 2.9) $\times 10^2$	PdBI
232.0	Jan. 25.18	4.5	(-1.6 $\pm$ 1.2) $\times 10^3$	PdBI
86.27	Feb. 1.47	4.5	(-5.0 $\pm$ 3.1) $\times 10^2$	PdBI
212.6	Feb. 1.47	4.5	(0.4 $\pm$ 1.3) $\times 10^3$	PdBI
86.27	Feb. 4.06	3.5	(-0.3 $\pm$ 4.8) $\times 10^2$	PdBI
231.5	Feb. 4.06	3.5	(-2.7 $\pm$ 4.3) $\times 10^3$	PdBI
86.64	Jan. 28.54	0.42	(-4.1 $\pm$ 9.1) $\times 10^3$	Pico Veleta
142.3	Jan. 28.54	0.42	(-3.7 $\pm$ 3.3) $\times 10^4$	Pico Veleta
228.9	Jan. 28.54	0.42	(-0.9 $\pm$ 1.9) $\times 10^4$	Pico Veleta
86.64	Jan. 30.53	1.42	(8.5 $\pm$ 3.5) $\times 10^3$	Pico Veleta
142.3	Jan. 30.53	1.42	(1.0 $\pm$ 2.9) $\times 10^4$	Pico Veleta
228.9	Jan. 30.53	1.42	(1.3 $\pm$ 1.2) $\times 10^4$	Pico Veleta
222	Jan. 24.68	0.65	(-4.1 $\pm$ 2.5) $\times 10^3$	JCMT*
222	Jan. 27.83	0.62	(0.7 $\pm$ 1.9) $\times 10^3$	JCMT*
353	Jan. 27.89	1.25	(-3.3 $\pm$ 1.2) $\times 10^3$	JCMT*
353	Jan. 29.88	1.0	(0.8 $\pm$ 1.5) $\times 10^3$	JCMT*
353	Feb. 4.83	0.7	(4.9 $\pm$ 1.5) $\times 10^3$	JCMT*
353	Feb. 5.85	1.5	(1.2 $\pm$ 1.1) $\times 10^3$	JCMT*
666	Jan. 27.88	1.0	(-0.2 $\pm$ 1.7) $\times 10^4$	JCMT*†
666	Jan. 29.87	0.75	(-0.8 $\pm$ 2.3) $\times 10^4$	JCMT*†
666	Feb. 4.82	0.5	(1.5 $\pm$ 1.8) $\times 10^4$	JCMT*†

Westerbork Synthesis Radio Telescope (WSRT): we used the Multi-Frequency Front Ends (MFFEs)²⁶ at 4.88 and 1.38 GHz, and the DCB continuum correlator, providing us with 8 bands of width 10 MHz at each frequency. The data were calibrated using interleaved observations of 3C48, 3C147 and 3C286 (ref. 27). VLA data are taken from ref. 19. IRAM Plateaux de Bure Interferometer (PdBI)²⁸: observations were done in five antenna configurations, on January 25 in 5B1 and on February 1, in 5B2. Spectral bandwidth of the dual frequency SIS receivers at 86.2 GHz was 560 MHz (LSB tuning), and 310 MHz (DSB tuning) at frequencies above 210 GHz. Good weather conditions allowed to use the atmospheric phase correction system to improve the signal-to-noise ratio. Results are given for position-fixed point source fits. Pico Veleta: observations with the IRAM 30-m telescope on the Pico Veleta in Spain²⁹ were done at 86.638, 142.33 and 228.930 GHz in wobbler switching mode with SIS receivers connected to the 1 GHz back-ends. Flux calibration was relative to W3 (OH)³⁰, using the gain-elevation curve from ref. 31. James Clerk Maxwell Telescope (JCMT): observations were made using SCUBA³² in the photometry mode, as for previous bursts³³. At 450/850 μm , GRB990123 was in the central pixel of the arrays (at 1,350 μm there is only one bolometer). The secondary was chopped between the source and sky at 7 Hz to take out small relative d.c. drifts between the bolometers and remove large-scale sky variations. The opacity at 450 and 850 μm was measured from skydips while using the continuously monitored 1.3-mm opacity as a guideline of any trends between the skydips. For the absolute flux calibration the gain was measured using Mars. The 3.3σ result at 353 GHz on February 4.76 UT is presumably a statistical fluctuation, as the result was not confirmed the following day.

* The 'integration' time always refers to the on-source plus off-source time. An 18-s integration thus amounts to 9-s on-source observation time.

† The last 450/850 μm measurement was excluded from the 450 reduction because of the low elevation of the source.

Table 3 Data on the eight comparison stars used

ID	Δ RA ($^{\circ}$)	Δ Dec. ($^{\circ}$)	<i>U</i>	<i>B</i>	<i>V</i>	<i>R</i>	<i>I</i>	<i>H</i>
1	−34.30	24.33	15.38 ± 0.01	15.45 ± 0.01	14.88 ± 0.01	14.57 ± 0.01	14.28 ± 0.01	13.54 ± 0.01
2	−92.37	53.94		17.84 ± 0.04	16.83 ± 0.01	16.00 ± 0.01	15.23 ± 0.02	
3	−96.92	48.55	16.88 ± 0.02	16.97 ± 0.02	16.34 ± 0.01	15.93 ± 0.01	15.57 ± 0.01	
4	49.55	30.08			20.34 ± 0.22	19.66 ± 0.37	18.07 ± 0.40	
5	−15.19	−29.00			20.76 ± 0.32	19.76 ± 0.51	19.09 ± 0.55	17.37 ± 0.03
6	23.94	−89.03		19.80 ± 0.23	18.95 ± 0.06	18.56 ± 0.10	18.23 ± 0.13	
7	−30.64	−74.52			19.86 ± 0.15	20.08 ± 0.32	20.09 ± 0.46	
8	64.13	−81.67	16.12 ± 0.01	16.18 ± 0.01	15.52 ± 0.01	15.14 ± 0.01	14.82 ± 0.01	

The magnitudes and offset from the OT of the eight comparison stars used. The error listed is the quadratic average of the measurement error (Poisson noise) and the error originating from measuring the OT with respect to several reference stars. Not included is the calibration error, which we estimate to be 0.10. The V-, R- and I-band images were calibrated using JKT images of the Landolt PG1528+062 and SA98 fields³⁴. The Landolt star SA104-335, imaged on Feb. 3.21 UT with the JKT, was used for the U-band calibration. We also calibrated the GRB field in B, V, R and I with images taken with the KPNO 0.9-m, around Feb. 29.07 UT. The V, R and I calibration agrees within 4% for the brighter reference stars with the JKT values. We adopt the V, R and I calibration of the JKT and the B calibration of the KPNO 0.9-m. For the H band, the standard star FS21 (ref. 35) was observed during the two nights the OT was detected.

is not a simple extrapolation of the GRB to much lower energies, but that there is a smooth connection between the optical emission measured 4.5 minutes after the burst and the afterglow detected later.

The reverse shock could cause emission^{3,4} that peaks in the optical waveband and is observed only during or just after the GRB, because the shock takes a time of the order of the duration of the burst to travel through the ejecta and then quickly stops emitting. The observed properties of GRB990123 appear to fit this model quite well. If this interpretation is correct, GRB990123 would be the first burst in which all three emitting regions have been seen: internal shocks causing the GRB, the reverse shock causing the prompt optical flash, and the forward shock causing the afterglow. The relatively smooth connection between the prompt optical emission and the afterglow fits the expectation that the reverse and forward shock start at about the same time. This model requires roughly equipartition magnetic fields in the ejecta, as does the prompt γ -ray emission.

The radio emission of GRB990123 is unique both due to its very early appearance and its rapid decline. A look at the broad-band

spectrum on January 24.65 (Fig. 4) reveals why this was so: it is a power law connecting the X-ray to optical wavebands, which peaks in the several tens of GHz range and turns over towards even lower frequencies. The synchrotron peak frequency is not very much above, or is even below, the radio waveband. This peak frequency is very different from that of GRB970508, for which the peak was still in the millimetre region after 12 days (ref. 12), and that of GRB971214, for which the peak was in the optical/near-infrared waveband after 0.6 days (ref. 23). We infer from this that the synchrotron peak frequency, ν_m , may span a large range in GRB afterglows. The rapid decline of the radio flux, caused by the low value of ν_m , may explain why some GRBs are not detected at radio wavelengths.

Regardless of the location of the peak, the steeply rising radio spectrum implies that the self-absorption frequency, ν_{as} , is ~ 30 GHz, similar to earlier bursts. In relativistic blast-wave models²⁴ the flux at the peak is constant during the blast-wave evolution, although the peak frequency decreases rapidly, as $t^{-3/2}$. In Fig. 4, two idealized extreme low-frequency extrapolations are shown. Both have a self-absorption frequency, ν_{as} , close to the radio waveband, but the lower

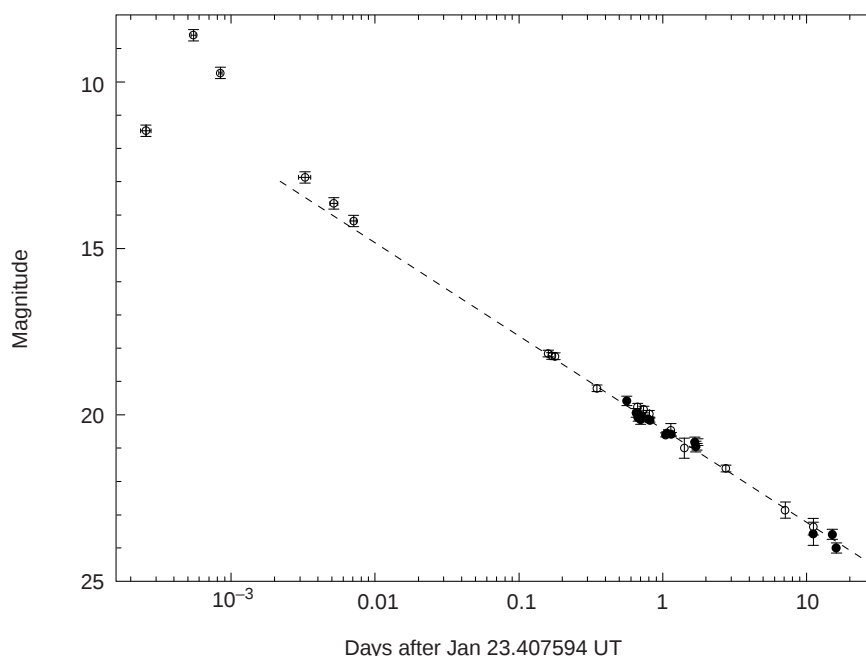


Figure 3 R-band light curve of the afterglow of GRB990123. The filled circles indicate results of our observations. The open circles are data taken from refs 37–43; we determined the corresponding magnitude for our calibration. The dashed line indicates a power-law fit to the light curve (for $t > 0.1$ days), which has exponent -1.12 ± 0.03 . Also included are the ROTSE data¹². ROTSE uses an

unfiltered CCD, but an equivalent V-band magnitude is reported; here we adopted a 0.1 mag error for the ROTSE data points and assuming a constant colour we have applied a colour correction of $V - R = 0.35 \pm 0.14$. The power-law fit is extrapolated backwards; it gives a good description of the last three ROTSE data points.

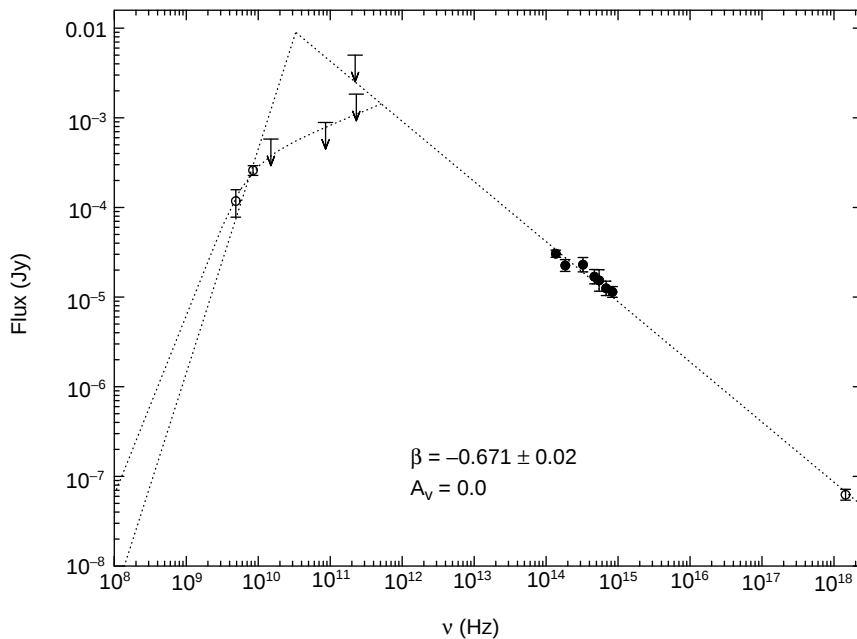


Figure 4 The spectral flux distribution of the afterglow at 1999 January 24.65 UT. Shown are the WSRT (see Table 1 for telescope names) 4.88 GHz detection (this work), the VLA 8.46 GHz detection⁴⁴, the KPNO 4-m U band, WIYN B, V, R and I, UKIRT H observations (this work), the K-band detection¹⁷, and the X-ray (2–10 keV) flux⁴⁵. Also shown are the Ryle 15 GHz, PdBI 86 and 230 GHz, and the JCMT 220 GHz 2σ upper limits (this work; we scaled these values back to the central time of the epoch by assuming that the source flux exhibits a $t^{-1.1}$ decay; that is, we used conservative limits). The photometric calibration has been taken from ref. 46 for B, V, R and I and from ref. 47 for H. We corrected the optical fluxes for Galactic foreground absorption ($A_V = 0.053$, as inferred from the dust maps of ref. 48). If more than one value per filter was available around the central time of the epoch, we took their weighted average. All values were brought to the same epoch by

applying a correction using the slope of the fitted light curve. We fitted the resulting optical to X-ray spectral flux distribution with a power law and an exponential optical extinction law, $F_\nu \propto \nu^\beta e^{-\tau}$, where we assume that the extinction optical depth, τ , is $\propto \nu$. The fit provides a negative extinction ($A_V < 0.19$; 90% confidence) and was subsequently fixed at zero. We find $\beta = -0.67 \pm 0.02$ ($A_V = 0$; $\chi^2_r = 6/7$); the resulting fit is indicated. Also shown for reference are two idealized extreme extrapolations of the low-frequency part of the spectrum: (1) a self-absorbed spectrum $F_\nu = F_{\nu_a}(\nu/\nu_a)^2(1 - \exp[-(\nu/\nu_a)^{-5/3}])$ (ref. 49), for a self-absorption frequency, $\nu_a = 7.5$ GHz and a flux density, $F_{\nu_a} = 350 \mu\text{Jy}$, and (2) the classic spectrum of a self-absorbed radio source with asymptotic spectral index $+5/2$. We note that realistic spectra are much rounder at the peak than the simple broken power-law spectra shown (see text for details).

one still has a synchrotron peak frequency, $\nu_m > \nu_a \sim 30$ GHz, so that there is an optically thin synchrotron regime, where the spectrum follows the standard low-frequency power law $F_\nu \propto \nu^{1/3}$. This implies that the radio flux should still rise after January 24.65, which is clearly inconsistent with the non-detections a few days after January 24.65. The second possibility is that $\nu_m < \nu_a$ already on January 24.65; then the flux at frequencies at and below ν_a would follow the optical decline if ν_a were constant. However, ν_a now begins to fall somewhat as well, and this will lessen the rate of decline at frequencies below ν_a (very much below ν_a it even rises). The shape of the spectrum will be much more rounded than the ideal cases shown, due to fluctuations in the field strength and Doppler broadening across the visible part of the shock, and the decline in radio is determined by the local spectral slope and the proximity of ν_a . Here we estimate that on January 24.65 the synchrotron peak frequency was at about 30 GHz, as was ν_a , causing a decline approximately as $t^{-0.9}$ of the fluxes detected by the Very Large Array (VLA) and the Westerbork Synthesis Radio Telescope (WSRT). The first non-detections by the VLA¹⁹ and WSRT after their respective detections are consistent with that decay rate.

The peak flux on January 24.65 (Fig. 4) is less than 2 mJy, while the last three ROTSE exposures have fluxes of ~ 10 mJy. If these exposures captured the early afterglow, the peak flux apparently decreased between 7.5 minutes and ~ 1.2 days after the burst, contrary to the prediction of a constant peak flux during the blast-wave evolution. There are, however, several possible explanations for the apparent decrease of the peak flux. If the peak frequency is less than the self-absorption frequency ν_a , then the peak flux will be lowered relative to its unabsorbed value. Or, part of the ROTSE flux could be due to a decaying prompt optical

component and thus the afterglow component at $t = 7.5$ minutes could be less than 10 mJy. Last, the afterglow peak flux may truly decay with time. The data, therefore, do not allow a strong distinction to be made between a constant and a decreasing peak flux of the afterglow spectrum.

A comparison of GRB990123 with other well-studied GRBs suggests some patterns. The cooling frequency of GRB970508, which separates quickly cooling synchrotron electrons from slowly cooling ones, was in the optical/near infrared region after 1.5 days (ref. 2). For GRB990123 and all other bursts in which the optical-to-X-ray afterglow spectrum could be reconstructed for the few days after the burst, any cooling break was located at or above X-ray frequencies. Both a higher cooling frequency and a lower peak frequency can be explained by a difference in one shock parameter: the magnetic field in the forward-shock region. The lower field also causes a lower peak flux^{24,25}, which, with the low peak frequency, conspires to make the radio emission weak and brief. The field energy density for the radio-quiet GRB971214 is estimated to be less than 10^{-5} times the equipartition value, and more than 1,000 times weaker than in GRB970508²⁵; the value for GRB990123 may be as low as 10^{-6} times equipartition. This suggests that we should distinguish between low-field afterglows, which are short and dim in radio emission, and high-field afterglows, which are bright and long-lived at radio wavelengths. The absence of detected radio emission in previous bursts with afterglows can then be explained by assuming they were low-field afterglows. We conjecture that such differences in field strength reflect differences in energy flow from the central engine; this flow could, for example take the form of a Poynting wind (an outflow dominated by electromagnetic waves). □

Received 4 February; accepted 3 March 1999.

1. Piran, T. Gamma-ray bursts and the fireball model. *Phys. Rep.* (in the press); preprint astro-ph/9810256 at (<http://xxx.lanl.gov>) (1998).
2. Galama, T. J. *et al.* The radio-to-X-ray spectrum of GRB970508 on 1997 May 21.0 UT. *Astrophys. J.* **500**, L97–L101 (1998).
3. Mészáros, P. & Rees, M. J. Optical and long-wavelength afterglow from gamma-ray bursts. *Astrophys. J.* **476**, 232–237 (1997).
4. Sari, R. & Piran, T. Predictions for the very early afterglow and the optical flash. Preprint astro-ph/9901338 at (<http://xxx.lanl.gov>) (1998).
5. Katz, J. I. Low-frequency spectra of gamma-ray bursts. *Astrophys. J.* **432**, L107–L109 (1994).
6. Costa, E. *et al.* Discovery of an X-ray afterglow associated with the γ -ray burst of 28 February 1997. *Nature* **387**, 783–785 (1997).
7. Feroci, M. *et al.* BeppoSAX follow-up search for the X-ray afterglow of GRB970508. *Astron. Astrophys.* **332**, L29–L32 (1998).
8. in't Zand, J. J. M. *et al.* Gamma-ray burst 980329 and its X-ray afterglow. *Astrophys. J.* **505**, L119–L122 (1998).
9. Nicastro, L. *et al.* BeppoSAX observations of GRB970402. *Astron. Astrophys.* **338**, L17–L20 (1998).
10. Piro, L. *et al.* Evidence for a late-time outburst of the X-ray afterglow of GRB970508 from BeppoSAX. *Astron. Astrophys.* **331**, L41–L44 (1998).
11. Connors, A. & Hueter, G. J. The X-ray characteristics of a classical gamma-ray burst and its afterglow. *Astrophys. J.* **501**, 307–324 (1998).
12. Akerlof, C. *et al.* Observation of contemporaneous optical radiation from a γ -ray burst. *Nature* **398**, 400–402 (1999).
13. Ford, L. A. *et al.* BATSE observations of gamma-ray burst spectra. 2: Peak energy evolution in bright, long bursts. *Astrophys. J.* **439**, 307–321 (1995).
14. Kouveliotou, C. *et al.* Identification of two classes of gamma-ray bursts. *Astrophys. J.* **413**, L101–L104 (1993).
15. Kelson, D. D. *et al.* *IAU Circ.* No. 7096 (1999).
16. Fruchter, A. *et al.* *GCN Circ.* No. 255 (1999).
17. Bloom, J. S. *et al.* *GCN Circ.* No. 240 (1999).
18. Bloom, J. S. *et al.* *GCN Circ.* No. 206 (1999).
19. Kulkarni, S. R. & Frail, D. A. *GCN Circ.* No. 239 (1999).
20. Fruchter, A. Was GRB980329 at $z \sim 5$? *Astrophys. J.* (in the press); preprint astro-ph/9810224 at (<http://xxx.lanl.gov>) (1998).
21. Waxman, E. Angular size and emission timescales of relativistic fireballs. *Astrophys. J.* **491**, L19–L22 (1997).
22. Wijers, R. A. M. J., Rees, M. J. & Mészáros, P. Shocked by GRB970228: the afterglow of a cosmological fireball. *Mon. Not. R. Astron. Soc.* **288**, L51–L56 (1997).
23. Ramaprakash, A. N. *et al.* The energetic afterglow of the γ -ray burst of 14 December 1997. *Nature* **393**, 43–46 (1998).
24. Sari, R., Piran, T. & Narayan, R. Spectra and light curves of gamma-ray burst afterglows. *Astrophys. J.* **497**, L17–L20 (1998).
25. Wijers, R. A. M. J. & Galama, T. J. Physical parameters of GRB970508 and GRB971214 from their afterglow synchrotron emission. *Astrophys. J.* (in the press); preprint astro-ph/9805341 at (<http://xxx.lanl.gov>) (1998).
26. De Bruyn, A. G. in *The Westerbork Observatory, Continuing Adventure in Radio Astronomy* (eds Raimond, E. & Genee, R.) 109–125 (ASSL Vol. 207, Kluwer Academic, 1996).
27. Baars, J. W. M., Genzel, R., Pauliny-Toth, I. I. K. & Witzel, A. The absolute spectrum of Cas A—An accurate flux density scale and a set of secondary calibrators. *Astron. Astrophys.* **61**, 99–106 (1977).
28. Guilloteau, S. *et al.* The IRAM interferometer on Plateau de Bure. *Astron. Astrophys.* **262**, 624–633 (1992).
29. Baars, J. W. M., Hooghoudt, B. G., Mezger, P. G. & De Jonge, M. J. The IRAM 30-m millimeter radio telescope on Pico Veleta, Spain. *Astron. Astrophys.* **175**, 319–326 (1987).
30. Reuter, H.-P. & Kramer, C. The mm-to-submm continuum spectra of W 3(OH) and K 3-50A. *Astron. Astrophys. Suppl.* **132**, 413–416 (1998).
31. Greve, A., Neri, R. & Sivers, A. The gain-elevation correction of the IRAM 30-m telescope. *Astron. Astrophys.* **339**, 183–186 (1998).
32. Holland, W. S. *et al.* SCUBA: A common-user submillimetre camera operating on the James Clerk Maxwell Telescope. *Mon. Not. R. Astron. Soc.* (in the press); preprint astro-ph/9809122 at (<http://xxx.lanl.gov>) (1999).
33. Smith, I. A. *et al.* SCUBA sub-millimeter observations of gamma-ray bursters I. GRB970508, 971214, 980326, 980329, 980519, 980703. *Astron. Astrophys.* (in the press); preprint astro-ph/9811026 at (<http://xxx.lanl.gov>) (1999).
34. Landolt, A. U. UBVR photometric standard stars in the magnitude range $11.5 < V < 16.0$ around the celestial equator. *Astron. J.* **104**, 340–376 (1992).
35. Casali, M. M. & Hawarden, T. G. *JCMT-UKIRT Newsl.* No. 3, 33 (1992).
36. Band, D. *et al.* BATSE observations of gamma-ray burst spectra. I—Spectral diversity. *Astrophys. J.* **413**, 281–292 (1993).
37. Gal, R. R., Odewahn, S. C., Bloom, S. J., Kulkarni, S. R. & Frail, D. A. *GCN Circ.* No. 207 (1999).
38. Garnavich, P., Jha, S., Stanek, K. & Garcia, M. *GCN Circ.* No. 215 (1999).
39. Masetti, N. *et al.* *GCN Circ.* No. 233 (1999).
40. Zhu, J., Chen, J. S. & Zhang, H. T. *GCN Circ.* No. 217 (1999).
41. Yadigaroglu, I. A., Halpern, J. P., Uglesich, R. & Kemp, J. *GCN Circ.* No. 242 (1999).
42. Yadigaroglu, I. A., Halpern, J. P., Uglesich, R. & Kemp, J. *GCN Circ.* No. 248 (1999).
43. Sokolov, V., Zharikov, S., Nicastro, L., Feroci, M. & Palazzi, E. *GCN Circ.* No. 209 (1999).
44. Frail, D. A. & Kulkarni, S. R. *GCN Circ.* No. 211 (1999).
45. Heise, J. *et al.* *IAU Circ.* No. 7099 (1999).
46. Bessell, M. S. UBVR photometry. II—The Cousins VRI system, its temperature and absolute flux calibration, and relevance for two-dimensional photometry. *Publ. Astron. Soc. Pacif.* **91**, 589–607 (1979).
47. Bessell, M. S. & Brett, J. M. JHKLM photometry—Standard systems, passbands, and intrinsic colors. *Publ. Astron. Soc. Pacif.* **100**, 1134–1151 (1988).
48. Schlegel, D. J., Finkbeiner, D. P. & Davis, M. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *Astrophys. J.* **500**, 525–553 (1998).
49. Granot, J., Piran, T. & Sari, R. Synchrotron self absorption in GRB afterglow. *Astrophys. J.* (in the press); preprint astro-ph/9808007 at (<http://xxx.lanl.gov>) (1999).

Acknowledgements. The United Kingdom Infra-Red Telescope is operated by the Joint Astronomy Centre on behalf of the UK Particle Physics and Astronomy Research Council (PPARC). Observations at the Wise Observatory are supported by the Basic Science Foundation of the Israeli Academy of Sciences. The JCMT is operated by the Joint Astronomy Centre on behalf of the UK PPARC, the Netherlands Organisation for Pure Research, and the National Research Council of Canada. We thank G. Watt, I. Robson, and J. van der Hulst for authorizing the JCMT observations. We also thank P. Meikle, P. Smith, D. Harmer and the KPNO GRB team for their observations, and G. Jacoby for providing BVRI photometric standards. T.J.G. is supported by NFRA, P.M.V. by the NWO Spinoza grant. C.K., D.L.B., K.H. and M.S.B. are supported by NASA.

Correspondence and requests for materials should be addressed to T.J.G. (e-mail: titus@astro.uva.nl).

Observation of contemporaneous optical radiation from a γ -ray burst

C. Akerlof*, R. Balsano†, S. Barthelmy‡§, J. Bloch†, P. Butterworth‡, D. Casperson†, T. Cline‡, S. Fletcher†, F. Frontera†, G. Gisler†, J. Heise‡, J. Hills†, R. Kehoe*, B. Lee*, S. Marshall*, T. McKay*, R. Miller†, L. Piro**, W. Priedhorsky†, J. Szymanski† & J. Wren†

* University of Michigan, Ann Arbor, Michigan 48109, USA

† Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

‡ NASA/Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

§ Universities Space Research Association, Seabrook, Maryland 20706, USA

|| Raytheon Systems, Lanham, Maryland 20706, USA

¶ Università degli Studi di Ferrara, Ferrara, Italy

Space Research Organization, Utrecht, The Netherlands

* Lawrence Livermore National Laboratory, Livermore, California 94550, USA

** Istituto Astrofisica Spaziale, Rome, Italy

The origin of γ -ray bursts (GRBs) has been enigmatic since their discovery¹. The situation improved dramatically in 1997, when the rapid availability of precise coordinates^{2,3} for the bursts allowed the detection of faint optical and radio afterglows—optical spectra thus obtained have demonstrated conclusively that the bursts occur at cosmological distances. But, despite efforts by several groups^{4–7}, optical detection has not hitherto been achieved during the brief duration of a burst. Here we report the detection of bright optical emission from GRB990123 while the burst was still in progress. Our observations begin 22 seconds after the onset of the burst and show an increase in brightness by a factor of 14 during the first 25 seconds; the brightness then declines by a factor of 100, at which point (700 seconds after the burst onset) it falls below our detection threshold. The redshift of this burst, $z \approx 1.6$ (refs 8, 9), implies a peak optical luminosity of $5 \times 10^{49} \text{ erg s}^{-1}$. Optical emission from γ -ray bursts has been generally thought to take place at the shock fronts generated by interaction of the primary energy source with the surrounding medium, where the γ -rays might also be produced. The lack of a significant change in the γ -ray light curve when the optical emission develops suggests that the γ -rays are not produced at the shock front, but closer to the site of the original explosion¹⁰.

The Robotic Optical Transient Search Experiment (ROTSE) is a programme optimized to search for optical radiation contemporaneous with the high-energy photons of a γ -ray burst. The basis for such observations is the BATSE detector on board the Compton Gamma-Ray Observatory. Via rapid processing of the telemetry data stream, the GRB Coordinates Network¹¹ (GCN) can supply estimated coordinates to distant observatories within a few seconds of the burst detection. The typical error of these coordinates is 5° . A successful imaging system must match this field of view to observe the true burst location with reasonable probability.

The detection reported here was performed with ROTSE-I, a two-by-two array of 35 mm camera telephoto lenses (Canon $f/1.8$, 200 mm focal length) coupled to large-format CCD (charge-coupled device) imagers (Thomson $14 \mu\text{m}$ $2,048 \times 2,048$ pixels). All four cameras are co-mounted on a single rapid-slewing platform capable of pointing to any part of the sky within 3 seconds. The cameras are angled with respect to each other, so that the composite field of view is $16^\circ \times 16^\circ$. This entire assembly is bolted to the roof of a communications enclosure that houses the computer control system. A motor-driven flip-away cover shields the detector from precipitation and direct sunlight. Weather sensors provide the vital information to shut down observations when storms appear,

augmented by additional logic to protect the instrument in case of power loss or computer failure. The apparatus is installed at Los Alamos National Laboratory in northern New Mexico.

Since March 1998, ROTSE-I has been active for $\sim 75\%$ of the total available nights, with most of the outage due to poor weather. During this period, ROTSE-I has responded to a total of 53 triggers. Of these, 26 are associated with GRBs and 13 are associated with soft γ -ray repeaters (SGRs). The median response time from the burst onset to start of the first exposure is 10 seconds.

During most of the night, ROTSE-I records a sequence of sky patrol images, mapping the entire visible sky with two pairs of exposures which reach a 5σ V magnitude threshold sensitivity (m_v) of 15. These data, approximately 8 gigabytes, are archived each night for later analysis. A GCN-provided trigger message interrupts any sequence in progress and initiates the slew to the estimated GRB location. A series of exposures with graduated times of 5, 75 and 200 seconds is then begun. Early in this sequence, the platform is 'jogged' by $\pm 8^\circ$ on each axis to obtain coverage of a four times larger field of view.

At 1999 January 23 09:46:56.12 UTC, an energetic burst triggered the BATSE detector. This message reached Los Alamos 4 seconds later and the first exposure began 6 seconds after this. Unfortunately, a software error prevented the data from being written to disk. The first analysable image was taken 22 seconds after the onset of the burst. The γ -ray light curve for GRB990123 was marked by an initial slow rise, so the BATSE trigger was based on relatively limited statistics. Thus the original GCN position estimate was displaced by 8.9° from subsequent localization, but the large ROTSE-I field of view was sufficient to contain the transient image. At 3.8 hours after the burst, the BeppoSAX satellite provided an X-ray localization¹² in which an optical afterglow was discovered by Odewahn *et al.*¹³ at Mt Palomar. This BeppoSAX position enabled rapid examination of a small region of the large ROTSE-I field. A bright and rapidly varying transient was found in the ROTSE images at right ascension (RA) 15 h 25 min 30.2 s, declination (dec.) $44^\circ 46' 0''$, in excellent agreement with the afterglow found by Odewahn *et al.* (RA 15 h 25 min 30.53 s, dec. $44^\circ 46' 0.5''$). Multiple absorption lines in the spectrum of the optical afterglow indicate a redshift of $z > 1.6$. Dark-subtracted and flattened ROTSE-I images of the GRB field are shown in Fig. 1. Details of the light curve are shown in Table 1.

By the time of the first exposure, the optical brightness of the transient had risen to $m_v = 11.7$ mag. The flux rose by a factor of 13.7 in the following 25 seconds and then began a rapid, apparently smooth, decline. This decline began precipitously, with a power-law slope of ~ -2.5 and gradually slowed to give a slope of ~ -1.5 . This decline, 10 minutes after the burst, agrees well with the power-law slope found hours later in early afterglow measurements¹⁴. These observations cover the transition from internal burst emission to external afterglow emission. The composite light curve is shown in Fig. 2.

Table 1 ROTSE-I observations

Exposure start	Exposure duration	Magnitude	Camera
–7,922.08	75	<14.8	C
22.18	5	11.70 ± 0.07	A
47.38	5	8.86 ± 0.02	A
72.67	5	9.97 ± 0.03	A
157.12	5	11.86 ± 0.13	C
281.40	75	13.07 ± 0.04	A
446.67	75	13.81 ± 0.07	A
611.94	75	14.28 ± 0.12	A
2,409.86	200	<15.6	A
5,133.98	800	<16.1	A

Exposure start times are listed in seconds, relative to the nominal BATSE trigger time (1999 January 23.407594 UT). Exposure durations are in seconds. Magnitudes are in the "V equivalent" system described in Fig. 1 legend. Errors include both statistical errors and systematic errors arising from zero-point calibration. They do not include errors due to variations in the unknown spectral slope of the emission. Magnitude limits are 5σ . The final limit results from co-adding the last four 200-s exposures and is quoted at the mean time of those exposures. Camera entries record the camera in which each observation was made.

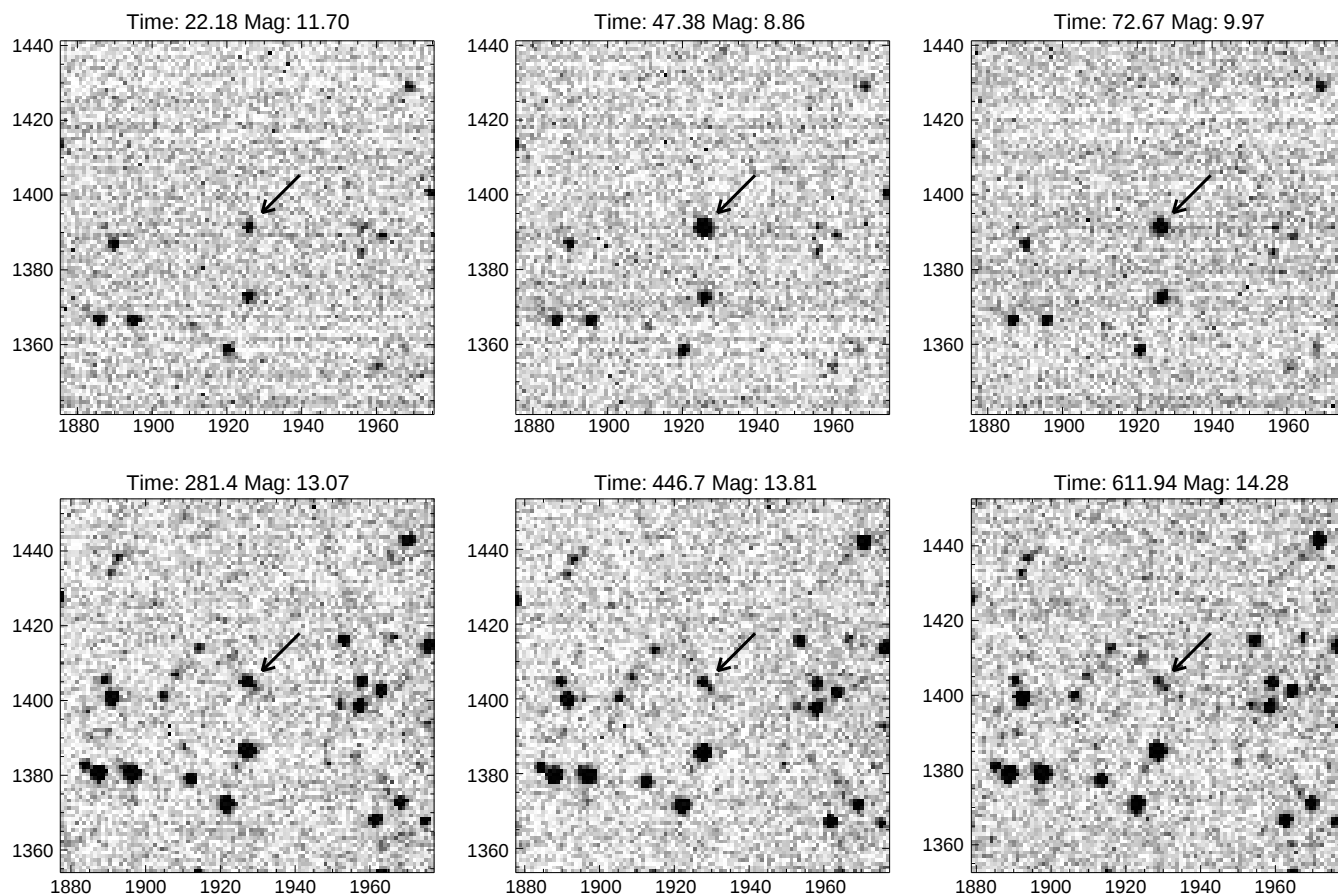


Figure 1 Time series images of the optical burst. Each image is $24''$ on a side, and represents 6×10^{-4} of the ROTSE-I field of view. The horizontal and vertical axes are the CCD pixel coordinates. The sensitivity variations are due to exposure time; the top three images are 5-s exposures, the bottom three are 75-s exposures. The optical transient (OT) is clearly detected in all images, and is indicated by the arrow. South is up, east is left. Thermal effects are removed from the images by subtracting an average dark exposure. Flat field images are generated by median averaging about 100 sky patrol (see text) images. Object catalogues are extracted from the images using SExtractor¹⁸. Astrometric and photometric calibrations are determined by comparison with the $\sim 1,000$ Tycho¹⁹ stars available in each image. Residuals for stars of magnitude 8.5–9.5 are $< 1.2''$. These images are obtained with unfiltered CCDs. The optics and CCD quantum efficiency limit our sensitivity to a wavelength range between 400 and 1,100 nm. Because this wide band is non-standard, we estimate a “V equivalent” magnitude by the following calibration scheme. For each Tycho star, a “predicted ROTSE magnitude” is compared to the

2.5 pixel aperture fluxes measured for these objects to obtain a global zero point for each ROTSE-I image. For the Tycho stars, the agreement between our predicted magnitude and the measured magnitude is ± 0.15 . These errors are dominated by colour variation. The zero points are determined to ± 0.02 . With large pixels, we must understand the effects of crowding. (This is especially true as we follow the transient to ever fainter magnitudes.) To check the effect of such crowding, we have compared the burst location to the locations of known objects from the USNO A V2.0 catalogue²⁰. The nearest object, $34''$ away, is a star with R-band magnitude $R = 19.2$. More important is an $R = 14.4$ star, $42''$ away. This object affects the measured magnitude of the OT only in our final detection. It can be seen in the final image to the lower right of the OT. A correction of $+0.15$ is applied to compensate for its presence. Magnitudes for the OT associated with GRB990123, measured as described, are listed in Table 1. Further information about the ROTSE-I observations is available at <http://www.umich.edu/~rotse>.

A number of arguments establish the association of our optical transient with the burst and the afterglow seen later. First, the statistical significance of the transient image exceeds 160σ at the peak. Second, the temporal correlation of the light curve with the GRB flux and the spatial correlations to the X-ray and afterglow positions argue strongly for a common origin. Third, the most recent previous sky patrol image was taken 130 minutes before the burst and no object is visible brighter than $m_v = 14.8$ mag. This is the most stringent limit on an optical precursor obtained to date. Searches further back in time (55 images dating to 28 September 1998) also find no signal. Finally, the ‘axis jogging’ protocol places the transient at different pixel locations within an image and even in different cameras throughout the exposure series, eliminating the possibility of a CCD defect or internal ‘ghost’ masquerading as a signal.

The fluence of GRB990123 was exceptionally high (99.6 percentile of BATSE triggers; M. Briggs, personal communication), imply-

ing that such bright optical transients may be rare. Models of early optical emission suggest that optical intensity scales with γ -ray fluence^{15–17}. If this is the case, ROTSE-I and similar instruments are sensitive to 50% of all GRBs. This translates to ~ 12 optically detected events per year. Our continuing analysis of less well-localized GRB data may therefore reveal similar transients. To date, this process has been hampered by the necessity of identifying and discarding typically 100,000 objects within the large field of view and optimizing a search strategy in the face of an unknown early time structure. The results we report here at least partially resolve the latter problem while increasing the incentive to complete a difficult analysis task. The ROTSE project is in the process of completing two 0.45-m telescopes capable of reaching 4 magnitudes deeper than ROTSE-I for the same duration exposures. If γ -ray emission in bursts is beamed but the optical emission is more isotropic, there may be many optical transients unassociated with detectable GRBs. These instruments will conduct sensitive searches

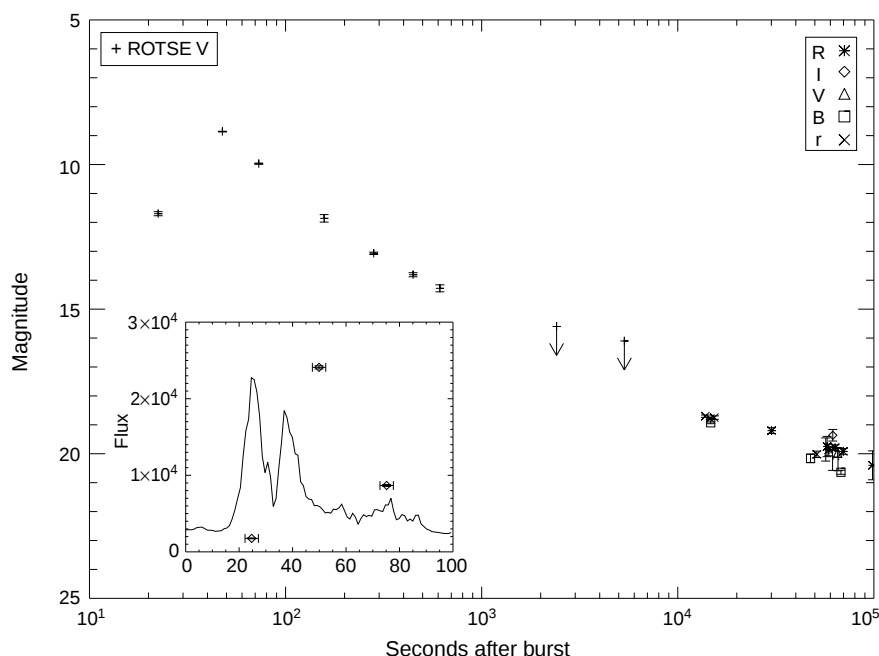


Figure 2 A combined optical light curve. Afterglow data points are drawn from the GCN archive^{21–33}. The early decay of the ROTSE-I light curve is not well fitted by a single power law. The final ROTSE limit is obtained by co-adding the final four 200-s images. The inset shows the first three ROTSE optical fluxes compared to the BATSE γ -ray light curve in the 100–300 keV energy band. The ROTSE-I fluxes

are in arbitrary units. Horizontal error bars indicate periods of active observation. We note that there is no information about the optical light curve outside these intervals. Vertical error bars represent flux uncertainties. Further information about GCN is available at <http://gcn.gsfc.nasa.gov/gcn>

for such events. We expect that ROTSE will be important in the exploration to come. □

Received 5 February; accepted 19 February 1999.

- Klebesadel, R. W., Strong, I. B. & Olson, R. A. Observations of gamma-ray bursts of cosmic origin. *Astrophys. J.* **182**, L85–L88 (1973).
- Piro, L. *et al.* The first X-ray localization of a γ -ray burst by BeppoSAX and its fast spectral evolution. *Astron. Astrophys.* **329**, 906–910 (1998).
- Costa, E. *et al.* Discovery of an X-ray afterglow associated with the γ -ray burst of 28 February 1997. *Nature* **387**, 783–785 (1997).
- Krimm, H. A., Vanderspek, R. K. & Ricker, G. R. Searches for optical counterparts of BATSE gamma-ray bursts with the Explosive Transient Camera. *Astron. Astrophys. Suppl.* **120**, 251–254 (1996).
- Hudec, R. & Soldán, J. Ground-based optical CCD experiments for GRB and optical transient detection. *Astrophys. Space Sci.* **231**, 311–314 (1995).
- Lee, B. *et al.* Results from Gamma-Ray Optical Counterpart Search Experiment: a real time search for gamma-ray burst optical counterparts. *Astrophys. J.* **482**, L125–L129 (1997).
- Park, H. S. *et al.* New constraints on simultaneous optical emission from gamma-ray bursts measured by the Livermore Optical Transient Imaging System experiment. *Astrophys. J.* **490**, L21–L24 (1997).
- Kelson, D. D., Illingworth, G. D., Franx, M., Magee, D. & van Dokkum, P. G. *IAU Circ.* No. 7096 (1999).
- Hjorth, J. *et al.* *GCN Circ.* No. 219 (1999).
- Fenimore, E. E., Ramirez-Ruiz, E., Wu, B. GRB990123: Evidence that the γ -rays come from a central engine. Preprint astro-ph/9902007 at (<http://xxx.lanl.gov>) (1999).
- Barthelmy, S. *et al.* in *Gamma-Ray Bursts: 4th Huntsville Symp.* (eds Meegan, C. A., Koskut, T. M. & Preece, R. D.) 99–103 (AIP Conf. Proc. 428, Am. Inst. Phys., College Park, 1997).
- Piro, L. *et al.* *GCN Circ.* No. 199 (1999).
- Odehahn, S. C. *et al.* *GCN Circ.* No. 201 (1999).
- Bloom, J. S. *et al.* *GCN Circ.* No. 208 (1999).
- Katz, J. I. Low-frequency spectra of gamma-ray bursts. *Astrophys. J.* **432**, L107–L109 (1994).
- Mészáros, P. & Rees, M. J. Optical and long-wavelength afterglow from gamma-ray bursts. *Astrophys. J.* **476**, 232–237 (1997).
- Sari, R. & Piran, T. The early afterglow. Preprint astro-ph/9901105 at (<http://xxx.lanl.gov>) (1999).
- Bertin, E. & Arnouts, S. SExtractor: Software for source extraction. *Astron. Astrophys. Suppl.* **117**, 393–404 (1996).
- Høg, E. *et al.* The Tycho reference catalogue. *Astron. Astrophys.* **335**, L65–L68 (1998).
- Monet, D. *et al.* *A Catalog of Astrometric Standards* (US Naval Observatory, Washington DC, 1998).
- Zhu, J. & Zhang, H. T. *GCN Circ.* No. 204 (1999).
- Bloom, J. S. *et al.* *GCN Circ.* No. 206 (1999).
- Gal, R. R. *et al.* *GCN Circ.* No. 207 (1999).
- Sokolov, V. *et al.* *GCN Circ.* No. 209 (1999).
- Ofek, E. & Leibowitz, E. M. *GCN Circ.* No. 210 (1999).
- Garnavich, P., Jha, S., Stanek, K. & Garcia, M. *GCN Circ.* No. 215 (1999).
- Zhu, J. *et al.* *GCN Circ.* No. 217 (1999).
- Bloom, J. S. *et al.* *GCN Circ.* No. 218 (1999).
- Maury, A., Boer, M. & Chaty, S. *GCN Circ.* No. 220 (1999).
- Zhu, J. *et al.* *GCN Circ.* No. 226 (1999).
- Sagar, R., Pandey, A. K., Yadav, R. K. R., Nilakshi & Mohan, V. *GCN Circ.* No. 227 (1999).
- Masetti, N. *et al.* *GCN Circ.* No. 233 (1999).
- Bloom, J. S. *et al.* *GCN Circ.* No. 240 (1999).

Acknowledgements. the ROTSE Collaboration thanks J. Fishman and the BATSE team for providing the data that enable the GCN localizations which made this experiment possible; and we thank the BeppoSAX team for rapid distribution of coordinates. This work was supported by NASA and the US DOE. The Los Alamos National Laboratory is operated by the University of California for the US Department of Energy (DOE). The work was performed in part under the auspices of the US DOE by Lawrence Livermore National Laboratory. BeppoSAX is a programme of the Italian Space Agency (ASI) with participation of the Dutch Space Agency (NIVR).

Correspondence and requests for materials should be addressed to C.A. (e-mail: akerlof@mich.physics.lsa.umich.edu).

A simple explanation of light emission in sonoluminescence

Sascha Hilgenfeldt*, Siegfried Grossmann† & Detlef Lohse‡

* Division of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, Massachusetts 02138, USA

† Fachbereich Physik der Universität Marburg, Renthof 6, D-35032 Marburg, Germany

‡ Department of Applied Physics and J. M. Burgers Centre for Fluid Dynamics, University of Twente, PO Box 217, NL-7500 AE Enschede, The Netherlands

Ultrasonically driven gas bubbles in liquids can emit intense bursts of light when they collapse¹. The physical mechanism for single-bubble sonoluminescence has been much debated^{2,3}. The conditions required for, and generated by, bubble collapse can be deduced within the framework of a hydrodynamic (Rayleigh–Plesset⁴) analysis of bubble dynamics and stability^{5,6}, and by considering the dissociation and outward diffusion of gases under the extreme conditions induced by collapse^{7,8}. We show here that by extending this hydrodynamic/chemical picture in a simple way, the light emission can be explained too. The additional elements that we add are a model for the volume dependence of the bubble's temperature^{9,10} and allowance for the small emissivity of a weakly ionized gas¹¹. Despite its simplicity, our approach can account quantitatively for the observed parameter dependences of

the light intensity and pulse width, as well as for the spectral shape and wavelength independence of the pulses^{12–15}.

The first reliable measurements of the pulse width of single-bubble sonoluminescence (SBSL) and its virtual independence on wavelength^{12–15} showed that simple Planck emission from a black body is not sufficient to explain the experimental observations, because it would result in much longer pulses at long wavelengths than in the short wavelength regime. Moss and collaborators¹¹ realized that the temperature-dependent photon absorption of the gas has to be taken into account to explain the deviation from black-body radiation. They found that the bubble is not an ideal absorber (that is, not a surface emitter like an ideal black body), but because of its tiny size most photons can escape without reabsorption and therefore the bubble is transparent for its own radiation most of the time (that is, it is a volume emitter). The photons carry information about the light production processes, and the radiation differs from black-body emission in both intensity and spectral shape.

To calculate the absorption coefficients (and thus the light emission) quantitatively, we need the temporal variation of the bubble radius $R(t)$ and of the gas temperature $T(t)$ inside the bubble. The bubble dynamics are well described by the Rayleigh-Plesset equation⁴ and the bubble temperature, which is assumed to be spatially uniform, follows directly from $R(t)$ via the near-adiabatic compression and heating of the collapsing bubble. (The assumption of uniform bubble temperature is supported by full numerical simulations of the bubble's interior^{11,16,17}.) To describe the heating, we use a polytropic exponent which itself varies with time, taken from Prosperetti's work^{9,10}. This leads to isothermal behaviour for most of the cycle, whereas adiabaticity is approached only near the radius minimum (see Section 1 of Supplementary Information). For the parameters of typical SBSL experiments, the maximum computed temperatures lie in a fairly narrow range of $T_{\max} \approx 20,000$ – $30,000$ K. (This is much higher than the $5,000$ K observed in multi-bubble sonoluminescence¹⁸.) All parameters entering the entire formalism described here are the appropriate material parameters for noble gases in water; no adjustable fitting parameters have been used. We consider only noble gases (especially argon and xenon), because all other (molecular) constituents of the gas dissociate and leave the bubble^{7,8}, as has been confirmed in experiment^{19,20}.

The computed maximum temperatures translate, via the Saha

equation²¹, into a rather small degree of ionization α . With ionization energies $E_{\text{ion}} \approx 16$ eV for argon and ≈ 12 eV for xenon, $\alpha \propto \exp(-E_{\text{ion}}/2k_B T)$ typically stays below a few per cent in Ar and below 10% in Xe. In this regime, three processes give the most important contributions to the photon absorption coefficient^{11,21}: free-free transitions of (1) electrons near ions (bremsstrahlung absorption) and (2) electrons near neutral atoms, and (3) bound-free transitions (ionization) (see Section 2 of Supplementary Information). We used generic formulae for the first two contributions and a hydrogen-like atom approximation, which has proved useful even for many-electron systems, for the third (ref. 21, and S.H., S.G. and D.L., manuscript in preparation). The total absorption coefficient (the inverse of the absorption length) at wavelength λ is obtained as the sum of the three terms, $\kappa_\lambda[T] = \kappa_\lambda^{(1)}[T] + \kappa_\lambda^{(2)}[T] + \kappa_\lambda^{(3)}[T]$. The emitted light power per wavelength interval (spectral radiance) follows from $R(t)$, $T(t)$, and $\kappa_\lambda[T(t)]$ (see Section 3 of Supplementary Information):

$$P_\lambda(t) = 4\pi R^2 I_\lambda^{\text{Pl}}[T(t)] \times \left(1 + \frac{\exp(-2\kappa_\lambda R(t))}{\kappa_\lambda R(t)} + \frac{\exp(-2\kappa_\lambda R(t)) - 1}{2\kappa_\lambda^2 (R(t))^2} \right) \quad (1)$$

Here, $I_\lambda^{\text{Pl}}[T(t)]$ is the Planck intensity emitted from a black body at temperature $T(t)$. The formula simplifies to the spectral radiance of a black body in the limit of $\kappa_\lambda R \rightarrow \infty$ and to that of a transparent emitter (the usual case for an SBSL bubble) with $P_\lambda \propto (\kappa_\lambda R) I_\lambda^{\text{Pl}}$ for small $\kappa_\lambda R$. P_λ inherits (via κ_λ) the sensitive dependence of $\alpha \propto \exp(-E_{\text{ion}}/2k_B T)$ on the temperature. This explains both the shortness of the pulse widths (the intensity is 'turned off' very quickly) and their approximate wavelength independence, as $\exp(-E_{\text{ion}}/2k_B T)$ does not vary with λ .

As photon absorption and emission must balance in local thermodynamic equilibrium (Kirchhoff's law), every contribution to the absorption corresponds to a light emission process. Sonoluminescence is therefore generated by thermal bremsstrahlung (the inverse process of (1), (2)) and recombination radiation (inverse of (3)).

Figure 1a shows examples for the spectral radiance following from equation (1) for strongly driven argon and xenon bubbles in water. They are in good agreement with experiment (not shown; compare ref. 3), and the stronger light emission for xenon (mainly

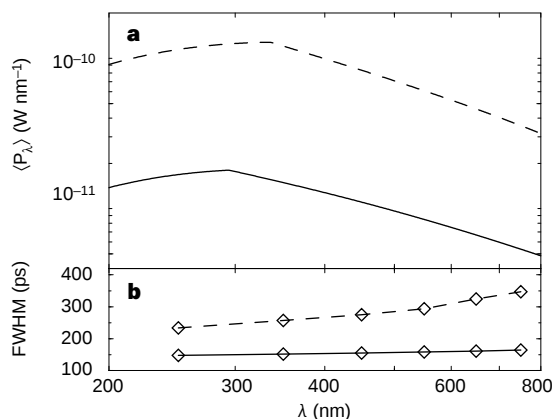


Figure 1 Spectral variation of intensity and pulse width of sonoluminescence pulses from xenon and argon bubbles. **a**, Spectral radiance from an argon (solid line) and a xenon (dashed) bubble, for $P_a = 1.3$ atm, $R_0 = 5.0$ μm . The intensity and shape of the experimentally reported spectra is reproduced very well (compare ref. 3), although the measured spectra decay slightly faster towards the red. **b**, The variation of pulse width with λ , computed for the total power contained in intervals of $\Delta\lambda = 100$ nm. This corresponds to experiments using filters of ~ 100 nm bandwidth¹².

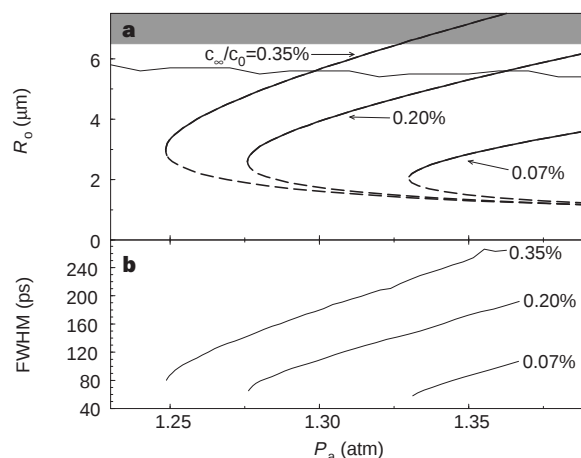


Figure 2 Size and light emission of diffusively stable argon bubbles. **a**, Curves $R_0(P_a)$ of stable (solid thick lines) and unstable (dashed) diffusive equilibria for $f = 20$ kHz at dissolved gas concentrations $c_\infty/c_0 = 0.07\%$, 0.20% and 0.35% (ref. 13), all given relative to the gas solubility c_0 . The parametric surface instability restricts R_0 to ~ 5.5 μm (calculated⁶, thin solid line) or ~ 7 μm (experimental²³, hatched region). **b**, Calculated pulse widths for the stable combinations of P_a and R_0 from **a**. The pulse widths and their dependence on P_a are in good agreement with experiments¹³.

due to its lower ionization energy) at equal parameters is reproduced. For both gases, a spectral maximum is predicted. It results from an absorption edge which reflects the detailed energy level structure of the noble gases (S.H., S.G. and D.L., manuscript in preparation). In reality, these maxima should lie at somewhat smaller λ as the level structure becomes modified at the high pressures inside the collapsed bubble. For xenon, the spectral maximum has been observed (at ~ 300 nm (ref. 3)); for argon, the experiments are not conclusive because of the light absorption of water which modifies the spectrum below $\lambda \approx 250$ nm. Figure 1b confirms that the full-width at half-maximum for the argon bubble shows practically no dependence on λ (as in experiment^{12,13,15}), whereas the present model predicts a quite pronounced variation of the largest and brightest xenon bubbles (though much smaller than in the black-body case). This variation should be detectable and awaits experimental verification.

A significant advantage of the present approach is the possibility to scan the whole parameter space of SBSL, because the calculations are very simple, and thus to reproduce experimentally observed parameter dependences of SBSL light emission. In experiment, the driving frequency f , the forcing pressure P_a , the water temperature T_w and the noble gas concentration c_∞ in the liquid far away from the bubble are the crucial adjustable parameters. To calculate $R(t)$, we need the bubble's ambient radius R_0 (radius of a stationary, undriven bubble) for given f , P_a , T_w and c_∞ . It has been shown^{6,22} that $R_0(f, P_a, T_w, c_\infty)$ is given by the condition of diffusive stability (dynamical equilibrium of gas exchange between the bubble interior and the surrounding liquid).

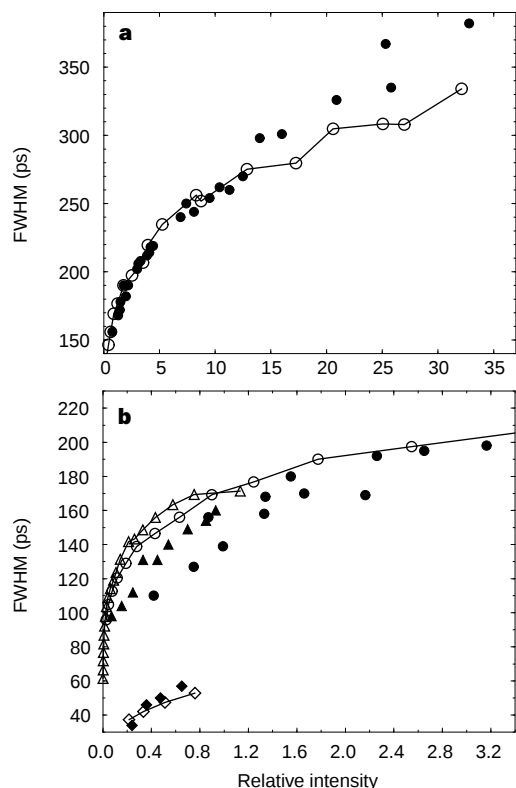


Figure 3 Calculated and measured sonoluminescence pulse widths as a function of light intensity. Open symbols denote theoretical results and filled symbols (from ref. 15) experimental data for $f = 34$ kHz. The larger experimental values for width and intensity are obtained for smaller water temperature, where the bubbles can be driven harder. **a**, Data for diffusively stable xenon bubbles at $c_\infty/c_0 = 0.4\%$. **b**, Diffusively stable Ar (triangles) and Xe (circles) bubbles, again for $c_\infty/c_0 = 0.4\%$. In addition, the values for $c_\infty/c_0 = 0.03\%$ argon are shown (diamonds), corresponding to the experiment on air at a partial pressure of 20 mm Hg (ref. 15).

Figure 2a shows the curves of diffusive equilibria in P_a – R_0 space (the branches with positive slope being stable) for several gas concentrations c_∞ used in experiment^{12,13}. As the experimenter changes P_a , the corresponding curve determines the response of the ambient size R_0 of the SBSL bubble, and calculating the light emission for these specific pairs (P_a, R_0) should therefore reproduce the measured photon numbers, pulse widths, and so on. The upper limits for P_a and R_0 are imposed by the onset of shape instabilities (compare Fig. 2a and refs 6, 23 and 24. Figure 2b shows the results for the widths (full-width at half-maximum) of the emitted SBSL pulses, as a function of P_a . The range of pulse widths is in excellent agreement with experiment. The shift in the P_a axis (~ 0.10 – 0.15 atm) with respect to the data in ref. 13 could be due to the simplified modelling, but also to experimental inaccuracies of at least 0.1 atm in measuring P_a (B. Gompf, personal communication).

A more stringent test of the theory presented here can be made by comparing its predictions to the extensive experimental data of Hiller *et al.*¹⁵ on spectral pulse widths and intensities, both of which are accurately measurable quantities. In ref. 15, the intensity values are given relative to the maximum intensity of an air bubble dissolved in water at 150 mm Hg partial pressure. This normalizing intensity was determined with the present model (the relevant partial pressure being only that of argon (that is 1.5 mm Hg), because of molecular dissociation), and used as the intensity unit for all theoretical results in Fig. 3. In Fig. 3a we compare for xenon bubbles the experimentally found dependence of FWHM on intensity (filled symbols) to our calculations (open symbols). The agreement is, despite the approximations involved in the model, excellent. Not only is the qualitative shape of the curve reproduced, but also a quantitative comparison shows that the pulse widths differ from the experimental values by no more than $\sim 5\%$ for moderate intensities, and by no more than $\sim 15\%$ for high intensities. Other features from experiment are reproduced in Fig. 3b for argon and xenon bubbles in the range of smaller intensities. We note first that both argon and xenon bubbles follow almost exactly the same line in this diagram, if the same noble-gas concentration is applied (here, $c_\infty/c_0 = 0.4\%$), although the values of P_a and R_0 are quite different at a given intensity. Second, the theory also reproduces the tiny pulse widths and intensities if air at 20 mm Hg partial pressure is used instead of Ar or Xe (Fig. 3b, bottom left), corresponding to a relative argon concentration of 0.03%. □

Received 25 June 1998; accepted 21 January 1999.

- Gaitan, D. F. *An Experimental Investigation of Acoustic Cavitation in Gaseous Liquids*. Thesis, Univ. Mississippi (1990).
- Crum, L. A. Sonoluminescence. *Phys. Today* **47**, 22–29 (1994).
- Barber, B. P., Hiller, R. A., Löfstedt, R., Putterman, S. J. & Weninger, K. R. Defining the unknowns of sonoluminescence. *Phys. Rep.* **281**, 65–143 (1997).
- Brennen, C. E. *Cavitation and Bubble Dynamics* (Oxford Univ. Press, 1995).
- Brenner, M. P., Lohse, D. & Dupont, T. Bubble shape oscillations and the onset of sonoluminescence. *Phys. Rev. Lett.* **75**, 954–957 (1995).
- Hilgenfeldt, S., Lohse, D. & Brenner, M. P. Phase diagrams for sonoluminescing bubbles. *Phys. Fluids* **8**, 2808–2825 (1996).
- Lohse, D., Brenner, M. P., Dupont, T., Hilgenfeldt, S. & Johnston, B. Sonoluminescing air bubbles rectify argon. *Phys. Rev. Lett.* **78**, 1359–1362 (1997).
- Lohse, D. & Hilgenfeldt, S. Inert gas accumulation in sonoluminescing bubbles. *J. Chem. Phys.* **107**, 6986–6997 (1997).
- Prosperetti, A. Thermal effects and damping mechanisms in the forced radial oscillations of gas bubbles in liquids. *J. Acoust. Soc. Am.* **61**, 17–27 (1977).
- Prosperetti, A. The thermal behaviour of oscillating gas bubbles. *J. Fluid Mech.* **222**, 587–616 (1991).
- Moss, W., Clarke, D. & Young, D. Calculated pulse widths and spectra of a single sonoluminescing bubble. *Science* **276**, 1398–1401 (1997).
- Gompf, B., Günther, R., Nick, G., Pecha, R. & Eisenmenger, W. Resolving sonoluminescence pulse width with time-correlated single photon counting. *Phys. Rev. Lett.* **79**, 1405–1408 (1997).
- Pecha, R., Gompf, B., Nick, G., Wang, Z. Q. & Eisenmenger, W. Resolving the sonoluminescence pulse shape with a streak camera. *Phys. Rev. Lett.* **81**, 717–720 (1998).
- Moran, M. J. & Sweider, D. Measurements of sonoluminescence temporal pulse shape. *Phys. Rev. Lett.* **80**, 4987–4990 (1998).
- Hiller, R. A., Putterman, S. J. & Weninger, K. R. Time-resolved spectra of sonoluminescence. *Phys. Rev. Lett.* **80**, 1090–1093 (1998).
- Vuong, V. Q. & Szeri, A. J. Sonoluminescence and diffusive transport. *Phys. Fluids* **8**, 2354–2364 (1996).
- Yuan, L., Cheng, H. Y., Chu, M. C. & Leung, P. T. Physical parameters affecting sonoluminescence: A self-consistent hydrodynamic study. *Phys. Rev. E* **77**, 4265–4280 (1998).
- Flint, E. B. & Suslick, K. S. The temperature of cavitation. *Nature* **253**, 1397–1399 (1991).
- Matula, T. J. & Crum, L. A. Evidence for gas exchange in single-bubble sonoluminescence. *Phys. Rev. Lett.* **80**, 865–868 (1998).

20. Ketterling, J. A. & Apfel, R. E. Experimental validation of the dissociation hypothesis for single bubble sonoluminescence. *Phys. Rev. Lett.* **81**, 4991–4994 (1998).
21. Zel'dovich, Y. B. & Raizer, Y. P. *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena* Vols I & II (Academic, New York, 1966).
22. Fyrrillas, M. M. & Szeri, A. J. Dissolution or growth of soluble spherical oscillating bubbles. *J. Fluid Mech.* **277**, 381–407 (1994).
23. Holt, G. & Gaitan, F. Observation of stability boundaries in the parameter space of single bubble sonoluminescence. *Phys. Rev. Lett.* **77**, 3791–3794 (1996).
24. Prosperetti, A. & Hao, Y. Modeling of spherical gas bubble oscillations and sonoluminescence. *Phil. Trans. R. Soc. Lond.* **357**, 203–224 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank S. Koehler, W. Moss and H. Stone for discussions. Support by the DFG and partial support by the NSF is acknowledged.

Correspondence and requests for materials should be addressed to D.L. (e-mail: lohse@tn.utwente.nl).

Immobility of protons in ice from 30 to 190 K

J. P. Cowin*, A. A. Tsekouras†, M. J. Iedema*, K. Wu* & G. B. Ellison‡

* Pacific Northwest National Laboratory, Box 999, MS K8-88, Richland, Washington 99352, USA

† Laboratory of Physical Chemistry, University of Athens, Athens GR-15771, Greece

‡ Department of Chemistry & Biochemistry, University of Colorado, Boulder, Colorado 80309, USA

The anomalously fast motion of hydronium ions (H_3O^+) in water is often attributed to the Grotthuss mechanism^{1,2}, whereby protons tunnel from one water molecule to the next. This tunnelling is relevant to proton motion through water in restricted geometries, such as in 'proton wires' in proteins³ and in stratospheric ice particles⁴. Transport of hydronium ions in ice is thought to be closely related to its transport in water^{1,2}. But whereas claims have been made that such tunnelling can persist even at 0 K in ice^{5–7}, counter-claims suggest that the activation energy for hydronium motion in ice is non-zero^{8–10}. Here we use 'soft-landing'^{11–13} of hydronium ions on the surface of ice to show that the ions do not seem to move at all at temperatures below 190 K. This implies not only that hydronium motion is an activated process, but also that

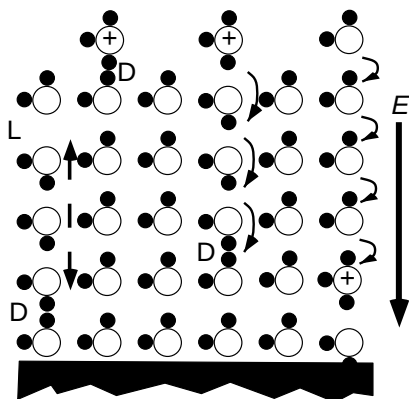


Figure 1 Ice structure and transport. A diagram of ice's hydrogen-bonding network is shown. On top were placed three hydronium ions (+). L and D hydrogen-bond defects are also present. At the top, 'soft-landed' hydronium ions form D defects. At right the hydronium ion's proton, in an electric field E , has hopped to an adjacent water four times, changing the hydronium's position. In the centre, a D defect moves away from the hydronium. At left, an L/D-defect pair formed and diffused apart.

it does not occur at anything like the rate expected from the Grotthuss mechanism. We also observe the motion of an important kind of defect in ice's hydrogen-bonded structure (the D defect). Extrapolation of our measurements to 0 K indicates that the defect is still mobile at this temperature, in an electric field of $1.6 \times 10^8 \text{ V m}^{-1}$.

Water's hydrogen-bonded network must be disrupted to solvate or transport ions. Figure 1 shows the hydrogen bonding in a thin film of water. Normally, each water molecule hydrogen-bonds with four neighbours. Three hydronium ions were initially placed on top of the film. At right, a proton has hopped four monolayers down, via the Grotthuss mechanism. Figure 1 shows D and L defects, which are misdirected hydrogen bonds, that put either two or zero protons between two adjacent oxygens instead of the usual one proton. These D/L defects facilitate the re-orientation of water molecules, creating water's high solvation power and dielectric constant ϵ .

The ions in Fig. 1 will move in their collectively self-generated electric field $|E|$. When all ions are at the top of a film (of thickness b) on an earthed substrate, this gives a 'film voltage' ΔV of

$$\Delta V = |E|b = \frac{Qb}{A\epsilon\epsilon_0} \quad (1)$$

where charge Q is deposited over area A , and ϵ_0 is the vacuum permittivity. The ion motion could be measured by the time-dependent current induced between an electrode placed on top of the film and the substrate, as in the classical 'proof' of proton tunnelling, the 95–180 K study by Eckener *et al.*⁵ They used 0.5-J laser pulses to "stimulate" the release of hydronium ions from hydrogen-gas-pretreated Pd powder. The powder was pressed onto one face of ice crystal wafers 90–300 μm thick; the opposing face was a biased electrode. However, these difficult experiments (never duplicated) are prone to artefacts^{10,14}. "Proton injection" experiments at much higher temperatures also claim a zero activation

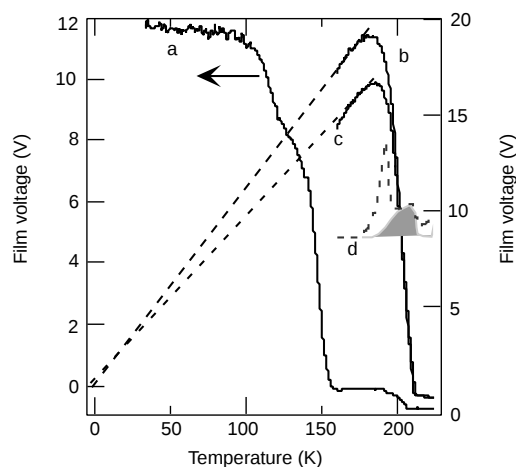


Figure 2 Lack of diffusion of hydronium ions in water ice. Data are shown for crystalline water films dosed with hydronium ions. Trace a, 2,880 monolayers (ML) grown at 150 K, ion-dosed (20 nC) at 33 K. Voltage (left axis) is shown as temperature T is increased at 0.167 K s^{-1} . Trace b, 5,980 ML grown at 165 K, then ion-dosed (700–3,800 nC) at 165 K. Trace c, 820 ML dosed at 165 K, then water dosed (480 ML) at 165 K. (Traces b, c use right axis.) Water desorption occurs near 200 K for these thick films (trace d, shown for 5,980 ML film). Maximum ion dose is 2.6% of a monolayer. The water desorption curve is double-peaked, as the Kelvin probe slows desorption from the region behind it. The shaded portion is the desorption peak to compare to the work-function changes. ϵ was measured to be about $2.3 \pm 33\%$ for the start of trace a, and $160 (\pm 50\%)$ for the start of curves b and c. Voltage remaining high from 33 to 150 K (trace a) and from 160 to 190 K shows that hydronium ions do not move in these temperature ranges. Dashed curve is the expected behaviour from equation (1), for ϵ proportional as $1/T$, if hydronium ions do not move.

barrier (ref. 6, and ref. 7 and refs therein). But like Eckener *et al.*, these higher temperature experiments probably produced mixtures of hydronium ions, counter-ions, D/L defects and even H_2 . Measurements using infrared spectroscopy and NMR indicate activation barriers of ~ 0.4 and 0.1 eV (refs 8, 9, 15), but distances were ill-defined, planar defects and traps were present, and counter-ions were produced.

To overcome these limitations, we gently “soft-landed”^{11–13,16} pure hydronium ions from our mass-selected 1-eV source onto ultrathin ice films. We do not use counter-ions; this is an advantage, as they interfere badly when determining ion mobility. We have used ‘soft-landed’ ions to probe ion diffusion in glassy hydrocarbon films¹⁷. These simpler systems allowed comparison with ion mobilities estimable from published viscosities; the good agreement validated our method (see Supplementary Information).

The film voltage ΔV in equation (1) changes the work function by $-\Delta V$, which is measurable by a non-contacting Kelvin probe¹⁸. If ϵ is roughly constant, then the work function measures the motion of the ions; b in equation (1) is replaced by the average ion height $\langle z \rangle$. ϵ varies slowly for ice at temperature $T > 160$ K. Then ϵ is “active” (that is, the time τ for reorientating water dipoles is ≤ 1 s), and equals about $(27,000 \text{ K})/T$ (ref. 1). Below 130 K $\tau \gg 1$ s, so ϵ then is constant at ~ 3.2 .

We prepared targets by depositing D_2O or H_2O films, 0.2 – $1.5 \mu\text{m}$ thick, on a Pt(111) single crystal, yielding single-crystal ice above 140 K (ref. 19). The ion beam energy is kept to 1 eV during ion deposition by adjusting the target voltage to compensate for the film voltage.

Figure 2 shows ΔV for D_3O^+ ions initially deposited on D_2O ice at 33 K. The initial film voltage is 11.7 V for 20 nC deposited ions. The calculated ϵ via equation (1) is $2.3 \pm 33\%$ assuming that the ions stay on top of the film. Significant ion motion would decrease the calculated ϵ , moving it unreasonably far from the established value of 3.2 (ref. 1). Thus hydronium ions do not move at 33 K, even in a $1.1 \times 10^7 \text{ V/m}$ field.

Figure 2 trace a shows as that, the temperature is ramped, the film voltage stays nearly constant up to 100 K, then drops in steps at 110, 150 and 200 K. Near 150 K ϵ should become “active”, increasing to over 100, accounting for this large voltage drop. Substantial water

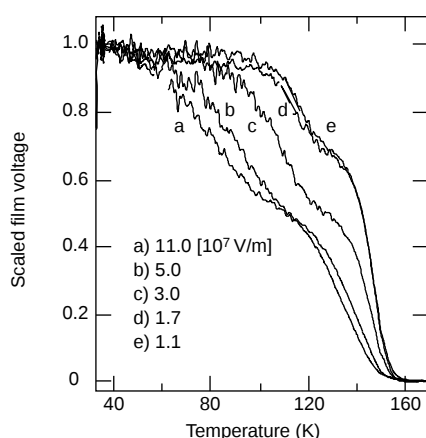


Figure 3 Evidence for hydrogen-bond defect motion. Shown is film voltage versus T for ions placed on crystalline water at various electric fields. The data has been scaled to the initial bias (with zero taken near 180 K to eliminate the 0.6 V work-function change from the first monolayer). Ice was grown at 150 K, ions dosed near 30 K. The rate of temperature increase is 0.167 K s^{-1} . Trace a, 600 ML D_2O , 24 V initial voltage; trace b, 1440 ML, 26 V; trace c, 2,780 ML, 30 V; trace d, 5,060 ML, 32 V; trace e, 2,880 ML, 12 V. Partial 1/3 to 1/2 voltage drop at lower temperatures than 150 K are due to D defects moving away from hydronium ions and then through the film. That trace e fits neatly into the family shows that there is a field effect at work, not just a thickness effect.

desorption begins around 180 K. The drop at 200 K is caused by desorption of the last of the water, and a work function change of $+0.6$ V. The lack of large voltage changes below 150 K shows that hydronium ions do not move from 33 to 120 K, and probably to 150 K.

To probe hydronium-ion motion above 150 K, we deposited substantially more ions, at temperatures between 140 and 165 K, where the dielectric constant is large. Data for H_2O ice grown at 165 K are given in Fig. 2 traces b and c. Equation (1) shows that, if the ions do not move, the voltage should vary as $1/\epsilon(T)$. For $T \geq 160$ K, ϵ varies as $1/T$, predicting a voltage proportional to T . Figure 2 shows the work function does increase linearly with T , until water desorption near 200 K. Thus hydronium ions do not move in ice from 33 to 190 K.

One concern is that there may be a surface trap or barrier preventing the ions from entering the ice bulk. We sandwiched ions inside crystalline ice films (Fig. 2 trace c), and saw no evidence of any surface trap: thus the non-mobility of the hydronium ions is a true bulk ice effect. The data in Fig. 2 are for D_3O^+ reported on both D_2O and H_2O . As exchange is facile above 160 K (refs 8, 9), no mobility occurs for either D_3O^+ in D_2O or H_3O^+ in H_2O .

The absence of the Grotthuss mechanism is easy to rationalize. Tunnelling over long distances should only occur if the potential is periodic. But normal ice is proton-disordered¹, thereby destroying the periodicity. Further, the hydronium ions should self-trap in a polarization well. A hydronium ion will force nearby water molecules to reorientate, pointing away their hydrogen ends. If the proton tries to move without the polarization pattern following it, it finds a deep well that it cannot escape. Solvation well effects have been surmised to occur elsewhere, as in calculated proton motions²⁰, or polaron trapping near -40°C (ref. 21).

Other experimenters^{8,9} have deduced that activated hydronium-ion motion in fact occurs at 140 or 115 K, rather than above 190 K as we find. We speculate that mobile D/L defects lower the activation barrier for hydronium–hydroxide ion-pair formation, permitting transient local ion pairs. This mimics hydronium-ion transport.

Figure 2 trace a shows a drop in ΔV near 110 K, where D/L defects should become mobile¹. The voltage always drops by about 0.3 to 0.5 times its initial value, as shown in Fig. 3 at various fields, and for a variety of other conditions (not shown). This is due to motion of D defects generated by the ions themselves: one D defect per hydronium ion. As in Fig. 1, each hydronium ion is expected to create D defects, owing to its extra proton and its weakly complexing oxygen end¹. Creation of a D/L defect pair requires¹ about 0.68 eV, so natural D/L defect concentrations should be insignificant below 150 K. The D/L defects have an effective charge of $\pm(0.38 e_e)$,

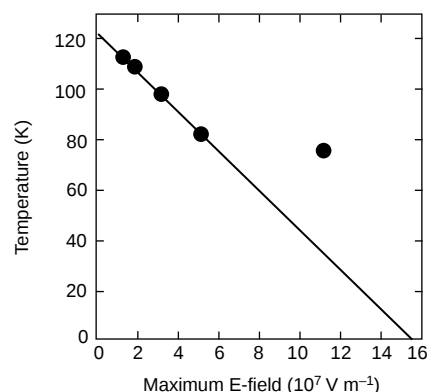


Figure 4 Hydrogen-bond defect motion versus electric field. Temperature for half-height of first voltage fall (hydrogen D-defect motion) of Fig. 3 data is plotted versus field strength. Linear fit has zero-field intercept of 124 K, and slope of $-8.1 \times 10^7 \text{ K V}^{-1} \text{ m}$.

where q_e is the electron charge. This is because their diffusion through a distance r 'flips' water dipoles over their entire path, creating a total dipole field change of $\pm(0.38 q_e)r$ (ref. 1). Thus a D defect reduces the field by 0.38 of the initial value as they traverse the film. Only D defects move; the field polarity keeps any negative-like L defects at the film top.

Figure 3 shows the shapes of the ΔV versus T curves for a variety of field strengths, $(1.1\text{--}11) \times 10^7 \text{ V m}^{-1}$. These fields are weak compared to electrochemical fields ($>10^9 \text{ V m}^{-1}$), and are comparable to those around ions in solution. The first voltage drop moves to much lower temperatures as the strength is increased: the D-defect drift velocity is not simply proportional to field strength. The field must lower the activation barrier for motion. The first-drop temperatures from Fig. 3 are plotted against field strength in Fig. 4; four points fall on a line. The fifth point corresponds to the highest field, which probably permitted ion motion during ion deposition. The zero-field limit is 124 K, in agreement with previous work for mobilization of pre-existing defects²². The D-defect mobility temperature falls to 0 K at an extrapolated field of $16 \times 10^7 \text{ V m}^{-1}$. This is the field at 17 Å from a singly charged ion in water, before solvation via reorientation occurs ($\epsilon = 3.2$). Thus an ion may be able to induce rapid D-defect migration as far away as 17 Å, greatly facilitating its solvation. Theoretical estimates of this field effect do not to our knowledge exist²³.

The work reported here suggests the need to refine the physical model of water ice. □

Received 19 October 1998; accepted 20 January 1999.

- Hobbs, P. V. *Ice Physics* (Clarendon, Oxford, 1974).
- Whalley, E., Jones, S. J. & Gold, L. W. (eds) *Physics and Chemistry of Ice* (Royal Soc. Canada, Ottawa, 1973).
- Pomes, R. & Roux, B. Structure and dynamics of a proton wire: A theoretical study of H^+ translocation along the single-file water chain in the gramicidin A channel. *Biophys. J.* **71**, 19–39 (1996).
- Bainco, R., Gertner, B. J. & Hynes, J. T. Proton transfer reactions at the surface of ice. Heterogeneous reactions involved in stratospheric ozone depletion. *Ber. Bunsenges. Phys. Chem.* **102**, 518–526 (1998).
- Eckener, U., Helmreich, D. & Engelhardt, H. in *Physics and Chemistry of Ice* (eds Whalley, E., Jones S. J. & Gold, L. W.) 242–245 (Royal Soc. Canada, Ottawa, 1973).
- Petrenko, V. F. & Maeno, N. Ice field transistor. *J. de Phys. C* **48**, 115–119 (1987).
- Petrenko, V. F. *Electrical Properties of Ice* (Spec. Rep. 93-20, US Army Cold Regions Research and Engineering Laboratory, Hanover, NH, 1993).
- Collier, W. B., Ritzhaupt, G. & Devlin, J. P. Spectroscopically evaluated rates and energies for proton transfer and Bjerrum defect migration in cubic ice. *J. Phys. Chem.* **88**, 363–368 (1984).
- Wooldridge, P. J. & Devlin, J. P. Proton trapping and defect energetics in ice from FT-IR monitoring of photoinduced isotopic exchange of isolated D_2O . *J. Chem. Phys.* **88**, 3086–3091 (1988).
- von Hippel, A., Runck, A. H. & Westphal, W. B. in *Physics and Chemistry of Ice* (eds Whalley, E., Jones, S. J. & Gold, L. W.) 236–241 (Royal Soc. Canada, Ottawa, 1973).
- Tsekouras, A. A., Iedema, M. J., Ellison, G. B. & Cowin, J. P. Soft-landed ions: a route to ionic solution studies. *Int. J. Mass Spectrom. Ion Proc.* **174**, 219–230 (1998).
- Kleyn, A. W. Ion-surface interactions—from channelling to soft-landing. *Science* **275**, 1440–1441 (1997).
- Miller, S. A., Luo, H., Pachuta, S. J. & Cooks, R. G. Soft-landing of polyatomic ions at fluorinated self-assembled monolayer surfaces. *Science* **275**, 1447–1450 (1997).
- Engelhardt, H. in *Physics and Chemistry of Ice* (eds Whalley, E., Jones, S. J. & Gold, L. W.) 226–235 (Royal Soc. Canada, Ottawa, 1973).
- Pfeifer, R. & Hertz, H. G. Activation energies of the proton-exchange reactions in water measured with the 1H-NMR spin echo technique. *Ber. Bunsenges. Phys. Chem.* **94**, 1349–1355 (1990).
- Strongin, D. R., Mowlem, J. K., Lynne, K. G. & Kong, Y. Efficient magnetic focusing of low-energy ions (1–5 eV) onto solids for use in surface chemistry studies. *Rev. Sci. Instrum.* **63**, 175–188 (1992).
- Tsekouras, A. A., Iedema, M. J. & Cowin, J. P. Soft-landed ion diffusion studies on vapor-deposited hexane films. *J. Phys. Chem.* (submitted).
- The KP5000 system (McAllister Technical Services, Coer d'Alene, Idaho 83815, USA).
- Materer, N. et al. Molecular surface structure of ice (0001): dynamical low-energy electron diffraction, total-energy calculations and molecular dynamics simulation. *Surf. Sci.* **381**, 190–210 (1997).
- Tuckerman, M., Laasonen, K., Sprik, M. & Parrinello, M. *Ab initio* molecular dynamics simulation of the solvation and transport of hydronium and hydroxyl ions in water. *J. Chem. Phys.* **103**, 150–161 (1995).
- Kunst, M. & Warman, J. M. Nanosecond time-resolved conductivity studies of pulse ionized ice. 2. The mobility and trapping of protons. *J. Phys. Chem.* **87**, 4093–4095 (1983).
- Johari, G. P. & Jones, S. J. Study of the low-temperature "transition" in ice 1 h by thermally stimulated depolarization measurements. *J. Chem. Phys.* **62**, 4213–4223 (1975).
- Hermansson, K. & Ojamae, L. On the role of electric fields for proton transfer in water. *Solid State Ionics* **77**, 34–42 (1995).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank S. E. Barlow for advice. This work was conducted in the Environmental Molecular Sciences Laboratory, a collaborative users' facility of the US Department of Energy (DOE), under the Office of Biological and Environmental Research, and was supported by the Division of Chemical Sciences, US DOE. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the US DOE.

Correspondence and requests for materials should be addressed to J.P.C. (e-mail: jp.cowin@pnl.gov).

Electrical conduction through DNA molecules

Hans-Werner Fink & Christian Schönenberger

Institute of Physics, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, Switzerland

The question of whether DNA is able to transport electrons has attracted much interest, particularly as this ability may play a role as a repair mechanism after radiation damage to the DNA helix¹. Experiments addressing DNA conductivity have involved a large number of DNA strands doped with intercalated donor and acceptor molecules, and the conductivity has been assessed from electron transfer rates as a function of the distance between the donor and acceptor sites^{2,3}. But the experimental results remain contradictory, as do theoretical predictions⁴. Here we report direct measurements of electrical current as a function of the potential applied across a few DNA molecules associated into single ropes at least 600 nm long, which indicate efficient conduction through the ropes. We find that the resistivity values derived from these measurements are comparable to those of conducting polymers, and indicate that DNA transports electrical current as efficiently as a good semiconductor. This property, and the fact that DNA molecules of specific composition ranging in length from just a few nucleotides to chains several tens of micrometres long can be routinely prepared, makes DNA ideally suited for the construction of mesoscopic electronic devices.

The physical meaning of conductivity implies that there are free charge carriers available that can move under ordinary thermal conditions. This quantity is not directly accessible from measurements of the transfer rates of injected hot electrons. The simplest

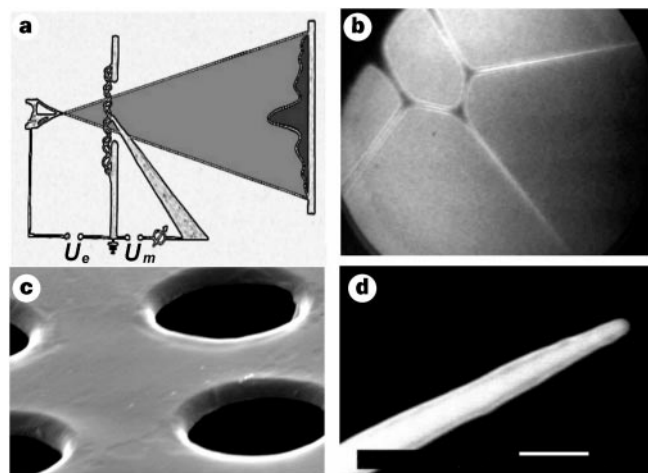


Figure 1 The low-energy electron point source (LEEPS) microscope used to investigate the conductivity of DNA. **a**, The atomic-size electron point source is placed close to a sample holder with holes spanned by DNA molecules. Due to the sharpness of the source and its closeness to the sample, a small voltage U_e (20–300 V) is sufficient to create a spherical low-energy electron wave. The projection image created by the low-energy electrons is observed at a distant detector. Between sample holder and detector, a manipulation-tip is incorporated (see text for details). This tip is placed at an electrical potential U_m with respect to the grounded sample holder and is used to mechanically and electrically manipulate the DNA ropes that are stretched over the holes in the sample holder. **b**, A projection image of λ -DNA ropes spanning a 2- μm -diameter hole. The kinetic energy of the imaging electrons is 70 eV. **c**, SEM image, showing the sample support with its 2- μm -diameter holes. **d**, SEM image of the end of a tungsten manipulation-tip used to contact the DNA ropes. Scale bar, 200 nm.

and most direct way to address the question of DNA conductivity is to arrange for a small potential difference between two well separated parts of the molecule and measure the resultant electrical current. Apart from the two contacts, the molecule should not be in contact with an object, but should instead be surrounded by vacuum to eliminate possible artefacts due to current paths other than those through the molecule itself. We have realized such an experiment by employing a modified low-energy electron point source (LEEPS) microscope⁵ (Fig. 1a). An earlier study involving mechanical and electrical manipulations of carbon nanotubes⁶ demonstrated the usefulness of the LEPS technology for probing elongated molecular objects, which are placed over holes in a suitable sample holder. An electron point source provides a low-energy coherent electron beam. Thus, in a simple projection set-up, we are able to image native DNA molecules. The low energy of the

imaging electrons, between 20 and 300 eV, provides high-contrast images and the absence of radiation damage⁷. The instrument is incorporated into a stainless-steel ultrahigh-vacuum chamber which provides an oil-free vacuum environment of $\sim 10^{-7}$ mbar. The λ -DNA molecules are placed onto a regular array of 2- μm holes in a carbon foil (Quantifoil, Jena, Germany). For a better electrical contact to the molecules, the sample holder has been covered with a gold layer. A scanning electron microscope (SEM) image of this sample holder is shown in Fig. 1c. The deposition of the molecules is achieved by first placing a drop of water containing $0.3 \mu\text{g ml}^{-1}$ of λ -DNA onto the sample holder, and then touching a piece of blotting paper to the rim of the drop to remove most of the liquid solution. (λ -DNA solution— $250 \mu\text{g ml}^{-1}$, with 10 mM Tris-HCl and 1 mM EDTA at pH 8.0 (Mannheim Boehringer)—was diluted by a factor of $\sim 1,000$ with deionized water, with a resistivity of $18 \text{ M}\Omega \text{ cm}$ at 25°C .) This preparation technique leads only very occasionally to individual DNA molecules spanning the holes of the sample holder: the λ -DNA usually forms networks within the holes (Fig. 1b), which implies that the DNA strands consist of a few molecules associated into a rope. If we use higher (lower) DNA concentration in solution, more (less) holes are covered by ropes while a solution containing no DNA results in no objects stretching over the holes. Due to the coherence of the electron point source, the projection images are in fact in-line holograms. In the rare case of observing a single DNA molecule stretching over a hole, the

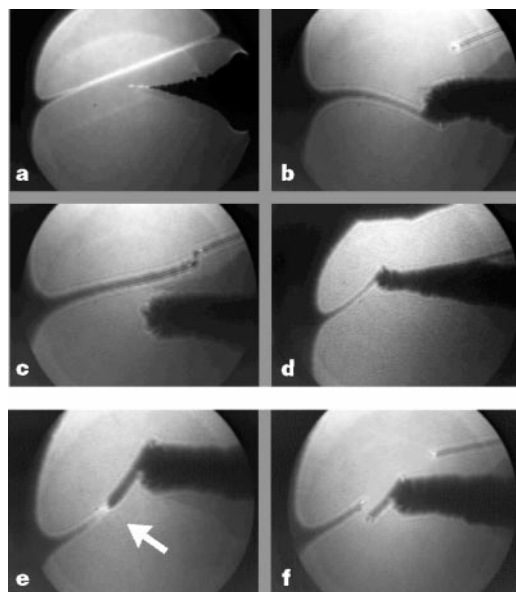


Figure 2 A sequence of LEPS images taken with 70-eV electrons, showing the mechanical and electrical manipulation of DNA ropes. **a**, A 2- μm hole in the sample holder is selected that is spanned by just one individual DNA rope. It exhibits a length of $\sim 1.5 \mu\text{m}$. The widening of the rope near the edge of the hole is probably due to capillary forces during evaporation of the solvent. At the surface the DNA molecules bind to the substrate and form a two-dimensional network; near the hole edges, a transition to the more compact three-dimensional rope spanning the hole is visible. Also apparent is the shadow image of the manipulation-tip that is imaged simultaneously, albeit at a low magnification as it is still at a macroscopic distance from the sample plane. Distances are measured by applying a step voltage to the piezo-ceramics on which the electron point source is mounted. This results in a lateral shift of the image corresponding to a distance given by the calibrated displacement of the piezo-ceramics. **b**, After the manipulation-tip has been moved onto the sample plane it is used to break the DNA rope into two parts. A 600-nm-long part attached to the left side of the sample holder has been bent down in order to probe its electrical conductivity by applying a voltage ramp to the manipulation-tip. Note that the magnification compared to the previous image is increased here, and only the very end of the longer (900-nm) DNA rope is apparent at the right part of the image. **c**, Further downwards motion of the manipulation-tip causes the 600-nm-long DNA piece to lose contact with the tip and flip back close to the position where it had originally been broken. **d**, An upward motion of the manipulation-tip results in a connection to both DNA ropes that now form two resistors in parallel. **e**, Extended bending and stretching of the 600-nm-long DNA rope results eventually in a defect. In contrast to the intact DNA rope, a voltage of -50 mV applied to the manipulation-tip manifests itself now (see pointer) in a potential drop along the DNA rope indicating poor conductivity at the defective region. **f**, Further mechanical stress applied to the defective DNA rope causes it to break, leaving a piece attached to the manipulation-tip.

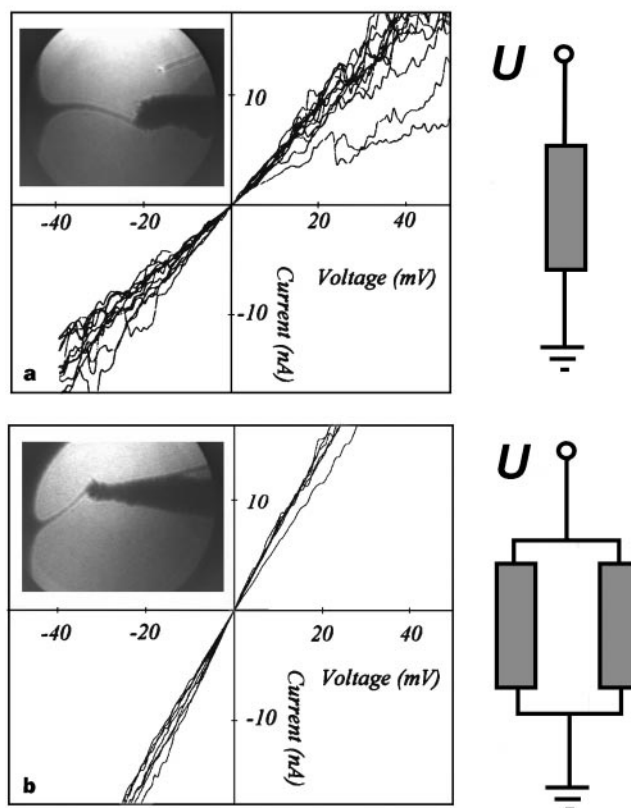


Figure 3 I - V characteristics of DNA ropes. **a**, I - V curve taken for a 600-nm-long DNA rope. In the range of $\pm 20 \text{ mV}$, the curves are linear; above this voltage, large fluctuations are apparent. From the linear dependence at low voltage we derive a resistance of about $2.5 \text{ M}\Omega$. **b**, I - V curve when the manipulation-tip is attached to both DNA ropes. The measured resistance drops to $1.4 \text{ M}\Omega$. The longer DNA rope is $\sim 900 \text{ nm}$ long, but due to the narrow angle it forms with the shank of the manipulation-tip, it is difficult to judge the actual position of the contact. Nevertheless, it appears that the situation can be viewed as a parallel connection of two resistances, $2.5 \text{ M}\Omega$ for the 600-nm rope and $3.3 \text{ M}\Omega$ for the 900-nm rope, accounting for the measured value.

reconstruction of the holograms reveals the shape of the 2-nm-wide molecule⁷. We do not observe charging effects while imaging the DNA, indicating that either the absorbed electrons are readily transported to the grounded sample holder or that no absorption of electrons takes place because all electrons are elastically scattered. We have also verified that the amount of inelastic scattered electrons is negligible. The lack of charging alone does not constitute evidence for the conductivity of the DNA. But some non-conducting objects, like small ionic crystals or latex spheres, show charging effects: the low-energy electrons are deflected by the electric fields associated with the charged objects.

Essential for the mechanical and electrical manipulation of DNA strands is the incorporation of an additional tip, the manipulation-tip (see SEM image, Fig. 1d), into the LEEPS microscope. With this additional tip, it is possible to achieve mechanical contact to a specific site of the DNA ropes, break the ropes at a certain distance from the rim of the hole in the sample holder, and place the ropes at a defined electrical potential difference with respect to the electrically grounded sample holder to which the other end of the molecules is fixed. These manipulations are performed *in situ* while the projection images of the molecules are observed on a TV monitor. In Fig. 2 we show a set of video frames illustrating the tasks necessary to probe the conductivity of DNA ropes. The circular field of view is that of the electron detector; it is not the boundary of the hole in the sample holder. First, the electron point source is moved along the hole array in the sample holder to search for a hole spanned by a single DNA rope. Once a suitable DNA rope has been identified, the manipulation-tip is moved into the field of view (Fig. 2a). The DNA molecules making up the rope and the manipulation-tip are then imaged simultaneously. As long as the manipulation-tip is still far away from the sample plane, it is imaged with a much lower magnification than the DNA rope. The tip is then moved into the sample plane to contact the DNA rope and break it into two parts. The manipulation-tip is then attached to the 600-nm-long part of the DNA rope which is forced to follow the downward motion of the tip (Fig. 2b). The conductivity of the 600-nm-long DNA rope is then probed by applying an electrical potential between the manipulation-tip and the sample holder, and monitoring the resulting current. Further downward motion of the manipulation-tip causes the DNA rope to lose contact with the manipulation-tip and to flip back close to the position where the rope had originally been broken (Fig. 2c). Next, an upward motion of the manipulation-tip causes a connection to both DNA ropes simultaneously (Fig. 2d). When an electrical potential is applied to the manipulation-tip, the resultant current can flow through either of the two DNA ropes. The projection images do not change while the applied potential is continuously varied between ± 50 mV, implying that there is no noticeable potential drop along the DNA rope. The sensitivity to small potential gradients is illustrated in Fig. 2e. The defective part of the DNA rope is clearly visible in this image, which was taken after repeated bending and stretching of the 600-nm-long DNA rope until it became defective. A negative potential applied to the manipulation-tip results in the part of the rope attached to the tip being 50 mV more negative than the grounded sample holder. This causes the low-energy electrons to be repelled, and so the projection of this part of the rope appears wider than the part attached to the grounded sample holder. With a positive voltage of the same magnitude the situation is reversed; the part attached to the sample holder appears wider. At zero bias, no effect on the image is visible. Further mechanical treatment by repeated motion of the manipulation-tip causes the DNA rope to eventually break at just this spot (Fig. 2f).

The apparent rigidity of the DNA ropes in vacuum might seem surprising, but appears to be in good agreement with standard models⁸ for the elastic behaviour of the DNA. The WLC ('worm-like chain') model treats the DNA as an elastic rod. The local elastic energy is proportional to the square of the curvature, which implies

that the lowest energy is associated with a rigid molecule. However, as the DNA is in solution, there will be an interplay between brownian motion and rigidity, which determines the persistence length over which the directionality of the molecule is maintained. In a vacuum environment, no brownian motion occurs, resulting in a minimal elastic energy for a molecule in a straight configuration. In a vacuum, the only external force acting on the DNA molecules is gravitation, which is extremely small for nanometre-sized objects.

We have measured current versus voltage (I - V) characteristics of the DNA ropes to obtain upper values for their resistivity. The voltage applied to the manipulation-tip is continuously swept and the current is monitored with an electrometer. These currents must be attributed to current through the molecules as they are in a vacuum: there is no alternative current path from the manipulation-tip to the grounded sample holder. Once the contact between manipulation-tip and the DNA rope is broken, the current drops to zero, confirming that no external artificial currents reach the electrometer. Also, the measured I - V characteristics do not depend on whether the situation is observed in the LEEPS microscope or not, indicating that the number of absorbed imaging electrons per second is below the noise level of the electrometer used to measure the current through the molecules. The I - V characteristics presented in Fig. 3 correspond to the situations shown in Fig. 2b and d. For the 600-nm-long DNA rope, we derive a resistance from the linear part of the I - V curve of 2.5 M Ω (Fig. 3a). If we assume that the DNA 'wire', which is made up of a few molecules, has a diameter of the same order of magnitude as that of a DNA double strand, (~ 2 nm), we arrive at a resistivity of the order of 1 m Ω cm. This value includes a contribution due to a finite contact resistance and therefore constitutes an upper limit; the resistivity attributed to the DNA rope alone will be smaller.

The I - V curve presented in Fig. 3b corresponds to the situation where the manipulation-tip is in contact with both DNA ropes, which act as resistors in parallel. The overall resistance drops to 1.4 M Ω , implying that the longer DNA rope has a resistance of 3.3 M Ω . Again, these values have to be considered as upper limits, due to the unknown contact resistances to the metallic reservoirs. The fact that we do not observe a potential drop along the DNA molecules, as discussed above, might possibly suggest that the electrical transport through the molecules is ballistic over large distances. The linearity of the measured I - V curves around the zero point implies that we observe an equilibrium property; that is, the measured current does not stem from the injection of hot electrons into the molecules.

Although our experiments establish that DNA molecules are molecular conductors, the mechanism allowing the transport of charges remains unclear. Ionic conduction can be ruled out as the water used to dissolve and subsequently deposit the DNA molecules evaporates or sublimates long before the vacuum environment (with a base pressure of $\sim 10^{-7}$ mbar) has been reached and the experiment has started. Moreover, an ionic conduction mechanism would require mobile ions, such as in a liquid or a molten phase, and macroscopic ion reservoirs to maintain the stable current we observe for as long as we keep the voltage applied. Neither of the two requirements is fulfilled in our experiments; there is no liquid water phase in the vacuum and no reservoirs supplying ions. Thus the intrinsic conduction mechanism must be of an electronic nature.

Although there is no solvent water left, the presence of tightly bound water, Tris-HCl and/or EDTA molecules (and possibly also counterions attached to the DNA) cannot be ruled out, even in a vacuum environment. While the electron transport mechanism must be of an electronic nature, any molecule attached to the DNA by an ionic or covalent bond might in principle affect the electronic structure, and hence the conductivity of the DNA molecules. Even if the inner π -electrons of the base stack are responsible for carrying the current, electric patch fields due to an 'outside' ionic bond may influence the π -electron states. There are

other factors that might influence DNA conductivity. For example, if vibrational modes of the molecules participate in the conduction mechanism, a certain length of the molecule may simply be required to allow for those collective effects, like solitons, to build up and to mediate electron transport. This would imply that restrictions of the degree of freedom of the DNA molecules such as through adsorption to a surface for example, could affect electrical conduction. There are probably a range of other issues that need to be considered to understand the details of the conduction mechanism on a molecular scale, but they do not affect our finding that DNA molecules, under the conditions described here, are electrical conductors. Further carefully designed experiments are now needed to explore the factors influencing DNA conductivity. Measurements of the temperature dependence of DNA resistivity, as well as the characterization of the electrical noise, should help to determine how DNA transports charges.

The detailed understanding of the conduction mechanism remains a challenge and, once achieved, should undoubtedly provide a better insight into the biological, chemical and physical properties of DNA molecules. The work reported here suggests that DNA molecules should be considered, among other candidates, as potential one-dimensional quantum wires for mesoscopic devices. □

Received 20 May 1998; accepted 20 January 1999.

1. Dandliker, P. J., Holmlin, R. E. & Barton, J. K. Oxidative thymine dimer repair in the DNA helix. *Science* **257**, 1465–1468 (1997).
2. Arkin, M. R. *et al.* Rates of DNA-mediated electron transfer between metallointercalators. *Science* **273**, 475–480 (1996).
3. Lewis, F. D. *et al.* Distance-dependent electron transfer in DNA hairpins. *Science* **277**, 673–676 (1997).
4. Beratan, D. N., Priyadarshy, S. & Risser, S. M. DNA: insulator or wire? *Chem. Biol.* **4**, 3–8 (1997).
5. Fink, H.-W., Stocker, W. & Schmid, H. Holography with low energy electrons. *Phys. Rev. Lett.* **65**, 1204–1206 (1990).
6. Schmid, H. & Fink, H.-W. Carbon nanotubes are coherent electron sources. *Appl. Phys. Lett.* **70**, 2679–2680 (1997).
7. Fink, H.-W., Schmid, H., Ermantraut, E. & Schulz, T. Electron holography of individual DNA molecules. *J. Opt. Soc. Am. A* **14**, 2168–2172 (1997).
8. Austin, R. H. Stretch genes. *Phys. Today* 32–38 (February 1997).

Acknowledgements. We thank our colleagues for discussions; G. Ehrlich for comments on the manuscript; and E. Ermantraut and K. Wohlfart for the design and production of the Quantifoil sample holders. This work was supported by the Swiss National Science Foundation and the Swiss priority program MINAST.

Correspondence and requests for materials should be addressed to H.-W.F. (e-mail: finkhw@ubach.unibas.ch).

Glacial–interglacial changes in ocean surface conditions in the Southern Hemisphere

F. Vimeux*, V. Masson*, J. Jouzel*, M. Stievenard* & J. R. Petit†

* Laboratoire des Sciences du Climat et de l'Environnement (UMR CEA/CNRS 1572), L'Orme des Merisiers, Bâtiment 709, CEA Saclay, 91191 Gif-sur-Yvette cedex, France

† Laboratoire de Glaciologie et de Géophysique de l'Environnement (CNRS), 54 rue Molière, Domaine Universitaire, BP 96, 38402 St-Martin d'Hères cedex, France

The stable-isotope signatures of oxygen and hydrogen in the water of preserved ice and snow are both widely used to infer local temperatures of past environments. A derived quantity based on these two signatures, the 'deuterium excess', provides additional palaeoclimatic information^{2–4}, as this parameter depends on the meteorological and oceanic characteristics of the water's source-regions (in particular, their temperature^{2,3} and relative humidity⁴). Published studies mainly focus on records from the past 40,000 years. Here we present a deuterium-excess history

obtained from ice cores from Vostok, East Antarctica, spanning the full glacial–interglacial cycle of the past 150,000 years. The deuterium-excess record shows a strong anticorrelation with the Earth's orbital obliquity (~41,000-year periodicity), and values are markedly higher during the cold stage 5d (following the last interglacial) than during the other cold stages. We interpret the relationship with obliquity as resulting from changes in the latitudinal insolation gradient affecting ocean surface conditions and, thus, the delivery of moisture to the polar region. We argue that the high 5d values, relative to other cold stages, are driven by relatively less moisture delivered from high latitudes, and more from low latitudes. The deuterium-excess in Antarctic precipitation thus provides long-term, spatially integrated information on ocean surface conditions and ocean/atmosphere circulations in the Southern Hemisphere.

The deuterium excess, d , (hereafter the excess) is defined as the deviation from the meteoric water line⁵: $d = \delta D - 8\delta^{18}O$ (see Fig. 1

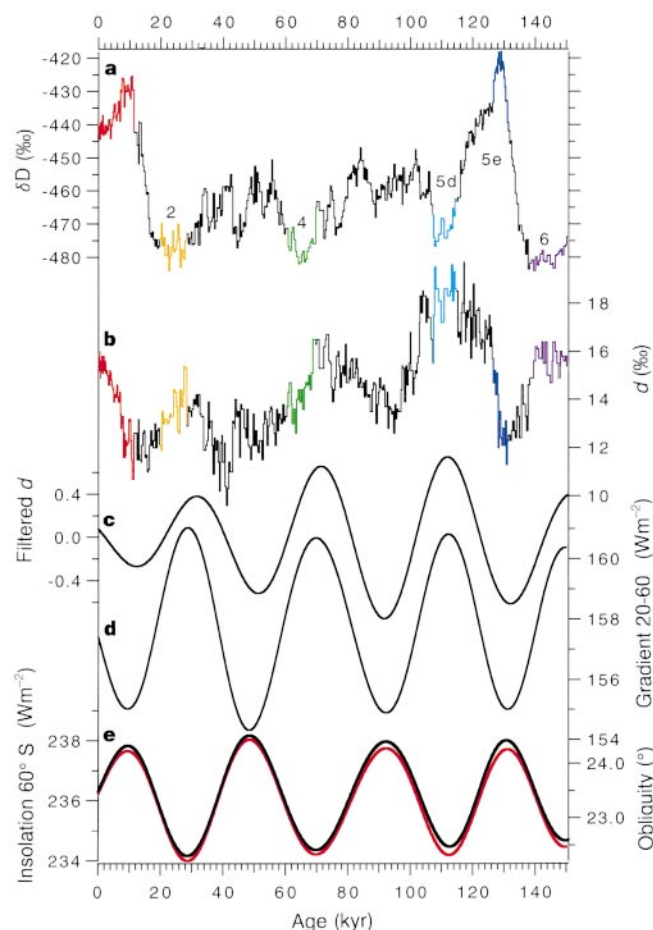


Figure 1 Time series of δD , deuterium excess, obliquity and insolation. **a**, δD ; **b**, deuterium excess, d . Measurements were performed along a shallow core from the surface to 138 m (BH8, drilled in 1996) and along deep core 3G from 138 to 2,083 m, with different sampling resolutions (50 cm between 0 and 138 m; 5 m down to 1413 m, 2 m below this¹⁰). Experimental accuracy for deuterium excess is $\pm 0.7\text{‰}$ ($\pm 0.5\text{‰}$ for δD and $\pm 0.05\text{‰}$ for $\delta^{18}O$) down to 1,413 m and $\pm 1.3\text{‰}$ below this ($\delta^{18}O$ measurements were performed with a dual mass spectrometer with a precision of $\pm 0.1\text{‰}$). Data are presented at a 5-m depth resolution representing a temporal resolution of 200 to 500 years. We use a modified version of the extended glaciological timescale, the accuracy of which is estimated to be ± 6 kyr (ref. 10). **c**, Gaussian-filtered deuterium excess in the obliquity band ($0.025 \pm 0.005 \text{ kyr}^{-1}$). **d**, Difference between mean annual insolation at 20°S and 60°S . **e**, Mean annual insolation at 60°S (black line), and obliquity¹⁴ (red line). (Here $\delta D = ((D/H)_{\text{sample}}/(D/H)_{\text{VSMOW}} - 1) \times 10^3$; $\delta^{18}O = ((^{18}O/^{16}O)_{\text{sample}}/(^{18}O/^{16}O)_{\text{VSMOW}} - 1) \times 10^3$.)

legend for nomenclature). The excess mainly reflects the kinetic fractionation occurring during non-equilibrium processes (evaporation above the ocean⁶ surface and snow formation⁷) due to the difference in diffusivity between heavy and light molecules. As shown by simple models based on the closure equation^{3,6}, the excess in the vapour above the ocean surface increases when the ocean surface temperature increases (+0.35 ‰ per °C) and decreases when relative humidity increases (−0.43 ‰ per %). Simple models^{3,8} have also shown that the initial excess value reflecting the oceanic source information is, at least partly, preserved all over the air-mass trajectory towards polar regions. Recent results obtained with a more complex atmospheric general circulation model (AGCM) calibrated with isotope tracer diagnostics support the idea that polar deuterium-excess values contain information on meteorological conditions at distant evaporative sources⁹.

The Vostok ice-core deuterium-excess profile is shown in Fig. 1 along with δD (ref. 10), a proxy of local past temperatures changes (see ref. 11 for a recent review). This record covers a full climatic cycle back to 150 kyr before present and considerably extends the 40-kyr Antarctic continuous record from Dome C (ref. 4; except for this site and for limited discontinuous Vostok data¹⁰, most studies of deuterium excess in Antarctica have focused on surface snow). Large variations are found in this record, from ~10 ‰ to 20 ‰. For the first 40 kyr, we note similar excess changes at Vostok and Dome C. In particular, for both records, the excess is 4–5 ‰ lower over the deglaciation than for the youngest part of the Holocene, indicating that glacial–interglacial fluctuations in Vostok excess represent at least a synoptic-scale pattern. We note a general anticorrelation between d and δD . However, Fig. 2, in which d is plotted as a function of δD , illustrates that the excess contains additional information compared to the deuterium profile. This is confirmed by a principal-component analysis performed on the two isotopes which provides (as second component) a combination similar to the excess profile.

During the two interglacial periods characterized by high δD values (Holocene in red and Eemian in dark blue; Figs 1 and 2) the excess increases with time. During transitions (characterized by large δD changes), two distinct types of excess behaviour appear; glacial inception with high excess values and deglaciations with low excess values. During glacial periods (low δD values), various excess levels can be present for a given temperature. In particular, stage 5d is characterized by the highest excess values (~19.5 ‰), whereas stages 2, 4 and 6 have lower excess values (~15 ‰). Before

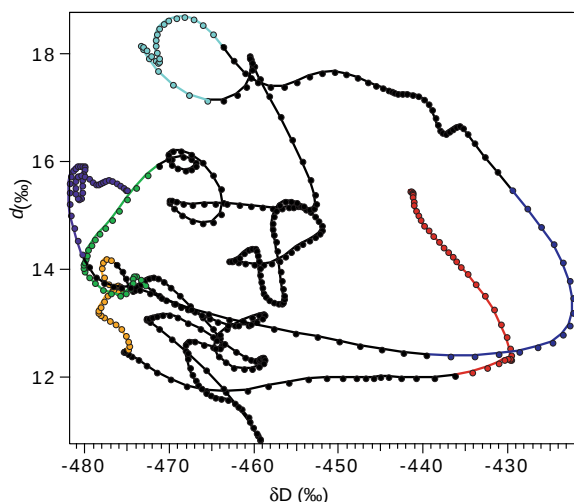


Figure 2 Deuterium excess versus δD (with smoothing over a 4,500-year running period). The colour represents the different stages as defined in Fig. 1: Holocene (red), Last Glacial Maximum (orange), stage 4 (green), stage 5d (light blue), last interglacial (dark blue) and previous glacial period (purple).

interpreting this obvious feature, we examine the spectral properties of the excess signal.

Various spectral methods have been used, and we have tested that the spectral properties are robust with respect to the empirically based definition of the excess by replacing 8.0 by any value between 7.0 and 9.0 (note that, in our record, the δD – $\delta^{18}O$ slope is 7.85). We conclude that only the obliquity (41-kyr periodicity) is strongly significant for the excess, and that it differs from the local δD temperature record for which both obliquity and precession (23-kyr) are significant (Fig. 3 shows only results from the multi-taper method¹²). A cross-spectral analysis of coherence and phase (using the Blackman–Tukey method¹³) between the excess and the obliquity points out their strong anticorrelation over the past 150 kyr. This is illustrated by comparing the excess values filtered in the 40-kyr band with the obliquity: the highest (lowest) values of deuterium excess are in phase with the lowest (highest) values of obliquity (Fig. 1). This orbital parameter strongly modulates the annual mean insolation¹⁴, in particular at high latitudes (Fig. 1) and we now examine how annual insolation changes could influence the deuterium excess of Antarctic snow.

First, Vostok local temperatures are partly influenced by the obliquity (as indicated by δD) and we could therefore think of the 40-kyr excess modulation as being due to local post-deposition effects such as those related to sublimation and depth-hoar formation (ice crystals formed by sublimation within snow, but beneath the snow surface) which, under certain conditions, may induce additional kinetic fractionation¹⁵. However, due to the extremely low temperatures in Vostok, this possibility can probably be discarded. Second, a recent study with an AGCM coupled with a slab ocean model has shown the effects of obliquity fluctuations on high-latitude climate due to the response of sea-ice cover. Increased obliquity (increased annual insolation at 60°S) and therefore reduced sea-ice cover¹⁶ induces a larger relative contribution of local cold moisture sources, resulting in low excess values, and vice

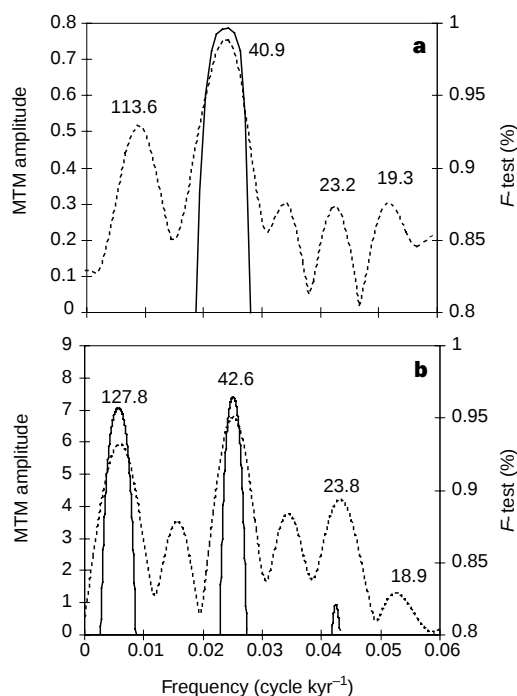


Figure 3 Harmonic analysis of the deuterium excess (a) and δD (b) interpolated at 0.5-kyr intervals. The analysis used the multi-taper method (MTM); the bandwidth parameter, $N\Omega$, is 4 and the number of tapers is 6 (see ref. 12 for details). This non-parametric method provides high-frequency resolution and has the advantage of providing a statistical test (solid line) for the validity of the amplitude (dashed line) and the location of each peak. Calculations have been performed with the AnalyseSeries software³⁰.

Table 1 Comparison of Vostok coldest stages 2, 4, 5d and 6

	Stage 2	Stage 4	Stage 5d	Stage 6	References
Vostok temperature (°C)*†	-6	-6	-5	-7	Ref. 10
Vostok deuterium excess (‰)*	14.5	15	18.5	15.5	This study
Vostok marine sodium (ng g ⁻¹)*	110	85	50	80	Ref. 22
Vostok dust (10 ⁻⁹ cm ³ g ⁻¹)*	690	510	4	404	Ref. 24
Sea surface temperature (°C) from MD 85668, 0° 01' S, 46° 02' E*†	-1.6	-0.7	+0.3	-0.4	Ref. 19
Sea surface temperature (°C) from MD 84 527, 43° 49' S, 51° 19' E*†	-4.0	-4.2	-4.1	-3.8	Ref. 29

* Averaged values during the coldest parts of stages 2, 4, 5d and 6.

† Temperatures deviations from their Holocene values.

versa. Although a significant influence of high-latitude oceans south of the polar front on central Antarctic precipitation is supported by AGCM simulations of present-day moisture sources, this contribution is too low (~15%) to explain by itself the 40-kyr modulation¹⁷.

Third, annual insulations are in phase with obliquity south of ~45° S and in antiphase from there to the Equator. Thus obliquity fluctuations significantly modulate the difference between annual insulations at low and high latitudes (~4% for the difference 20° S–60° S, Fig. 1) and have effects on both atmospheric and ocean circulations (moisture and heat transport). A minimum latitudinal insolation gradient weakens the atmospheric meridional circulation and reduces the contribution of low-latitude moisture to Antarctic precipitation. In parallel, less oceanic heat transport from low to high latitudes results in lower sea surface temperature (SST) at high latitudes. Both effects contribute to a colder origin for the precipitation and therefore excess values are lower when the insolation gradient is weak, that is, when the obliquity is large. This change in source temperatures also affects the Antarctic snow isotopic content: if relative humidity is kept constant, a colder source implies an increase of δD and $\delta^{18}O$ and vice versa. Indeed this may explain the generally observed anticorrelation between d and δD . Part of the 40-kyr modulation in the Vostok isotopic record may result from an associated shift of the polar front which is itself reflected in the deuterium excess (by inducing a parallel shift in the source areas).

Although the influence of changes in relative humidity cannot be discarded⁴, we favour the SST interpretation for two reasons. First, simple isotopic models show that the imprint of the source temperature on Antarctic snow deuterium excess increases inland while the imprint of the source relative humidity⁸ decreases. Second, we have analysed the modern and glacial climates simulated by various AGCMs within the Paleoclimate Modeling Intercomparison Project¹⁸. All models show only weak relative humidity changes (typically 1%) above southern oceans north of the polar front.

The changes in respective contributions of low- and high-latitude moisture also provide a plausible explanation for the extremely high excess level during cold stage 5d compared with the other cold periods (stages 2, 4, 6). The high latitudes are characterized by similarly cold temperatures during all these stages (Table 1), but, during glacial inception, the southern tropics are warmer than during stages 2, 4 and 6 as generally shown by available ocean cores^{19–21} covering these periods. Indeed, the high excess level implies a strong contribution of moisture from the warm ocean and suggests that, during glacial inception, low-latitude evaporation was maintained at its interglacial level.

Moreover, the Vostok marine sodium profile²² shows much lower concentrations during stage 5d than during stages 2, 4 and 6 (Table 1), probably associated with a decreased atmospheric transport from production regions (high latitudes) to the ice cap²³. Dust concentrations in the Vostok ice are also different in stage 5d (no dust) than in the other glacial periods²⁴. The absence of dust peaks during stage 5d may be also due to a weaker atmospheric circulation^{23,24} over the dust sources (mainly South America²⁵).

The combined information from marine sodium and dust suggests a weakened atmospheric circulation at high latitudes²⁴ during stage 5d and therefore less atmospheric transport of high latitude moisture. Thus, two phenomena are different at stage 5d

compared with stages 2, 4 and 6: warmer SST in the equatorial region, possibly maintaining local evaporation, and less atmospheric transport from the high latitudes as indicated by dust and marine sodium records. As a result, the relative contribution of low latitudes to Vostok precipitation would be higher during stage 5d than during other Vostok cold stages, explaining higher values of deuterium excess.

The study of the last climatic cycle in Vostok points out the importance of the moisture origin, and the relative contribution of low and high latitudes, in controlling the excess values. Therefore, the deuterium excess provides integrated, and thus unique, information on oceanic conditions and their past fluctuations, complementing relatively scarce southern deep-sea core data. The use of isotopic AGCMs should help to quantify better the changes in the deuterium excess and the link of relevant oceanic surface parameters with obliquity. Here, using the deuterium excess, we have pointed out the effect of the insolation gradient on the relative contribution of low and high latitudes to Antarctic precipitation (this relative contribution is inaccessible by other means). Indeed, changes in deuterium excess at Vostok also provide an indirect measure of southern latitudinal moisture transport above the ocean. The importance of this latitudinal moisture transport on thermohaline oceanic circulation is now fully recognized for the North Atlantic (see ref. 26 and references therein). This should also hold true for the Southern Ocean. In turn, obtaining information on latitudinal moisture transport may help the understanding of the role of the ocean in linking the climates of the Southern and Northern hemispheres²⁷.

The deuterium excess at Vostok also suggests that oceanic conditions during glacial stage 5d were more similar to an interglacial than to other glacial stages (2, 4, 6); this indicates a significant difference in the hydrological cycle between a cold stage just following an interglacial peak (5e) and a full glacial cold stage. Preliminary results¹⁰ seem to indicate a similar situation for the previous glacial inception (stage 7). The possibility now offered of obtaining a continuous profile of deuterium excess over four climatic cycles in Vostok²⁸ should allow further investigation of this difference in hydrological cycle; it should also allow further study of the link between deuterium excess, insolation, and surface-condition changes. □

Received 27 February 1998; accepted 18 January 1999.

- Dansgaard, W. Stable isotopes in precipitation. *Tellus* **16**, 436–447 (1964).
- Dansgaard, W., White, J. W. & Johnsen, S. J. The abrupt termination of the Younger Dryas climate event. *Nature* **339**, 532–534 (1989).
- Johnsen, S. J., Dansgaard, W. & White, J. W. C. The origin of Arctic precipitation under present and glacial conditions. *Tellus B* **41**, 452–468 (1989).
- Jouzel, J., Merlivat, L. & Lorius, C. Deuterium excess in an East Antarctic ice core suggests higher relative humidity at the oceanic surface during the last glacial maximum. *Nature* **299**, 688–691 (1982).
- Craig, H. Isotopic variations in meteoric waters. *Science* **133**, 1702–1703 (1961).
- Merlivat, L. & Jouzel, J. Global climatic interpretation of the deuterium-oxygen 18 relationship for precipitation. *J. Geophys. Res.* **84**, 5029–5033 (1979).
- Jouzel, J. & Merlivat, L. Deuterium and oxygen 18 in precipitation: modelling of the isotopic effects during snow formation. *J. Geophys. Res.* **89**, 11749–11757 (1984).
- Petit, J. R., White, J. W. C., Young, N. W., Jouzel, J. & Korotkevich, Y. S. Deuterium excess in recent Antarctic snow. *J. Geophys. Res.* **96**, 5113–5122 (1991).
- Armengaud, A., Koster, R., Jouzel, J. & Ciais, P. Deuterium excess in Greenland snow: Analysis with simple and complex models. *J. Geophys. Res.* **103**, 8947–8953 (1998).
- Jouzel, J. et al. Climatic interpretation of the recently extended Vostok ice records. *Clim. Dyn.* **12**, 513–521 (1996).
- Jouzel, J. et al. Validity of the temperature reconstruction from ice cores. *J. Geophys. Res.* **102**, 26471–26487 (1997).
- Thomson, D. J. Spectrum estimation and harmonic analysis. *Proc. IEEE* **70**, 1055–1096 (1982).

13. Jenkins, G. M. & Watts, D. G. *Spectral Analysis and its Applications* (Holden-Day, San Francisco, 1968).
14. Berger, A. L. Long-term variations of daily insolation and quaternary climatic change. *J. Atmos. Sci.* **35**, 2362–2367 (1978).
15. Satake, H. & Kawada, K. The quantitative evaluation of sublimation and the estimation of original hydrogen and oxygen isotope ratios of a firn core at East Queen Maud Land, Antarctica. *Bull. Glacier Res.* **15**, 93–97 (1997).
16. Gallimore, R. G. & Kutzbach, J. E. Snow cover and sea ice sensitivity to generic changes in Earth orbital parameters. *J. Geophys. Res.* **100**, 1103–1120 (1995).
17. Delaygue, G., Masson, V., Jouzel, J., Koster, R. D. & Healy, R. J. The origin of Antarctic precipitation: a modelling approach. *Tellus* (submitted).
18. Joussame, S. & Taylor, K. E. in *Proc. 1st AMIP Sci. Conf.*, WCRP Vol. 92 425–430 (World Climate Research Program, Monterey, CA, 1995).
19. Bard, E., Rostek, F. & Sonzogni, C. Interhemispheric synchrony of the last deglaciation inferred from alkenone palaeothermometry. *Nature* **385**, 707–710 (1997).
20. Schneider, R. R., Müller, P. J. & Rühland, G. Late Quaternary surface circulation in the east equatorial South Atlantic: Evidence from alkenone sea surface temperatures. *Paleoceanography* **10**, 197–219 (1995).
21. Van Campo, E., Duplessy, J. C., Prell, W. L., Barratt, N. & Sabatier, R. Comparison of terrestrial and marine temperature estimate for the past 135 kyr off South East Africa: a test for GCM simulations of palaeoclimate. *Nature* **348**, 209–212 (1990).
22. Legrand, M. R., Lorius, C., Barkov, N. I. & Petrov, V. N. Vostok (Antarctica) ice core: atmospheric chemistry changes over the last climatic cycle (160000 years). *Atmos. Environ.* **22**, 317–331 (1988).
23. Genthon, C. Simulations of desert dust and sea-salt aerosols in Antarctica with a general circulation model of the atmosphere. *Tellus B* **44**, 371–389 (1992).
24. Petit, J. R. *et al.* Paleoclimatological and chronological implications of the Vostok core dust record. *Nature* **343**, 56–58 (1990).
25. Basile, I. *et al.* Patagonian origin of glacial dust deposited in East Antarctica (Vostok and Dome C) during glacial stages 2, 4 and 6. *Earth Planet. Sci. Lett.* **143**, 573–590 (1997).
26. Stocker, T. S. & Schmittner, A. Influence of CO₂ emission rates on the stability of the thermohaline circulation. *Nature* **388**, 862–865 (1997).
27. Blunier, T. *et al.* Asynchrony of Antarctica and Greenland climate during the last glacial. *Nature* **394**, 739–743 (1998).
28. Petit, J. R. *et al.* Four climate cycles in Vostok ice core. *Nature* **387**, 359–360 (1997).
29. Pichon, J. J. *et al.* Surface water temperature changes in the high latitudes of the southern hemisphere over the last glacial-interglacial cycle. *Paleoceanography* **7**, 289–318 (1992).
30. Paillard, D., Labeyrie, L. & Yiou, P. Macintosh program performs time-series analysis. *Eos* **77**, 379 (1996).

Acknowledgements. Vostok is a joint project between Russia, France and USA. We thank all Russian, French and US participants in drilling, field work and ice sampling. We acknowledge the Russian Antarctic Expeditions (RAE), the Mining Institute, the Institut Français de Recherches et Technologies Polaires (IFRTP) and the Division of Polar Programs (NSF) for the logistic support. We thank D. Paillard and P. Yiou for their fruitful comments and S. Johnsen for his constructive criticism. This work project is supported by PNEDC (Programme National d'études de la Dynamique du Climat) and by the CEC (Commission of European Communities) Environment Programme.

Correspondence and requests for materials should be addressed to F.V. (e-mail: vimeux@lsce.saclay.cea.fr)

Gravity-driven continental overflow and Archaean tectonics

R. C. Bailey

Geology and Physics Departments, University of Toronto, 60 St George Street, Toronto, Ontario M5S 1A7, Canada

Whether modern tectonic processes differ substantially from those in Archaean times (>2,500 Myr ago) remains controversial. One view¹ is that Archaean tectonic processes were some combination of modern ones, occurring faster or more shallowly because of the larger heat output of the early Earth, but others² have proposed that significantly different processes operated. Here I argue that gravitational spreading of Archaean continents would have caused them continuously and pervasively to 'overflow' onto adjacent ocean basins, and that this process would have naturally ceased at the end of the Archaean era. Because modern continental crust is believed to be ductile rather than brittle below a depth corresponding to a temperature of about 350–400 °C (ref. 3), it seems likely that such a ductile zone was universally present within the hotter Archaean continental crust. If the mean geothermal gradient of the continents had exceeded ~25–30 °C km⁻¹, then the resulting ductile zone would have caused continental overflow to occur, and such a process can account for many of the distinctive peculiarities observed in the Archaean geological record. The cessation of continental overflow corresponds naturally to the stabilizing 'cratonization' which marked the end of the Archaean era, with its timing dependent on the evolution of

both the geothermal gradient in the continents and the depth of the ocean basins.

Processes such as tectonic compression and subduction- or plume-related magmatism can thicken continental crust. The amount of such thickening is limited by gravitational collapse⁴, which thins and spreads continental crust. Gravitational spreading uses the gravitational energy of overthickened crust to provide the power needed to drive extensional normal faulting in the upper brittle crust, viscous simple shear of the lower ductile crust, and viscous shear in the mantle as mantle flow accommodates isostatic adjustments of the crust. If the crustal ductile channel is sufficiently inviscid, significant upper-crustal topographic relief cannot be maintained in the spreading region, and must be concentrated in a 'front' at the leading edge (or edges) of a spreading plateau, as observed in Tibet today. Such a front can take one of two forms: if the crust into which the front is propagating also possesses a ductile channel, spreading will drive lower-crustal material into this channel and thicken it, 'inflating' the crust in the manner modelled by Bird⁵. Alternatively, and this is the case I analyse here, the adjacent crust may contain no ductile channel: in particular, it may be oceanic crust, which, although being as hot as the advancing felsic material, is not ductile because of its mafic composition. In this case, frontal advance can proceed only if continental material overflows the oceanic crust by thrusting, as shown schematically in Fig. 1.

To assess the degree of crustal overthickening required to drive an overflow front, I estimate here the magnitude of the various energy dissipation terms relative to the available gravitational power. Because viscous shear heating must be quadratic in the overflow front velocity, it must be negligible at the onset of overflow relative to the other components of the energy budget, which must be linear in the overflow front velocity; the threshold for overflow is therefore determined by the balance of these linear terms. As I will show, the threshold is primarily determined by the balance of gravitational input power and normal-fault dissipation; all other terms are much less important. To show the relative importance of these terms, it is convenient to have a specific and simple example in mind; here I evaluate them for the case of a microcontinent of width L and much longer strike length whose brittle upper crust and deep crustal ductile zone are of uniform thicknesses b and h , respectively (Fig. 1). Spreading is assumed to be perpendicular to strike, increasing L with frontal velocity v_f . For simplicity, I assume that the frontal geometry does not change with extension, although the thicknesses b and h must decrease proportionally with the rate of increase of L in order to conserve mass, the former by normal-fault rotation in the brittle crust, and the latter by Couette (simple shear) flow in the ductile zone (implying that the flow velocity tends to zero at the base of the ductile zone). The simplifying assumption that the depth of the brittle–ductile transition is approximately independent of strain rate is reasonable; the strongly thermally activated nature of any reasonable crustal rheology means that the effective viscosity profile

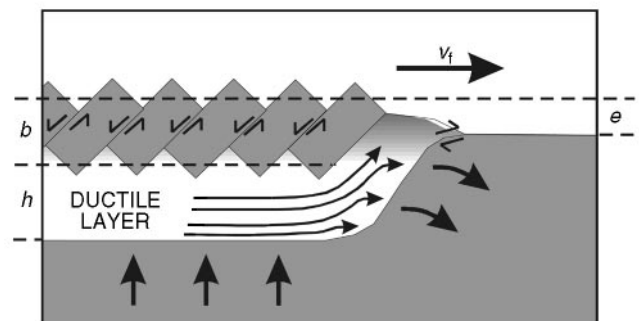


Figure 1 Cross-section perpendicular to strike of a simplified Archaean microcontinent. Symbols are defined in the text.

should not shift up or down more than about 2 km, even for an order-of-magnitude change in strain rate.

The power source is the gravitational potential energy of the microcontinent. The gravitational energy per unit area of a lithospheric column (relative to the purely mafic lithosphere of the ocean basin) is^{6–8}

$$u_g = \int z \Delta \rho(z) g dz \quad (1)$$

where $\Delta \rho(z)$ is the density difference from the oceanic reference column and g is the acceleration of gravity. For the simplified crustal model consisting of isostatically compensated submarine continental crust of density ρ_c under water of density ρ_w over mafic lithosphere of density ρ_m , u_g is evaluated as $(1/2)g\rho_c e^2$, e is crustal elevation above the adjacent ocean floor, and ρ_e is an 'effective' density given by $(\rho_c - \rho_w) + (\rho_c - \rho_w)^2/(\rho_m - \rho_c)$. The total energy U_g , per unit strike length of the microcontinent, is $(1/2)Lg\rho_e e^2$; the rate at which extension extracts power from this is obtained by differentiating U_g with respect to time, noting that conservation of mass requires that eL remain constant, and that the frontal velocity v_f is defined as dL/dt . The analysis yields the gravitational power output as $P_g = (0.5)\rho_e g e^2 v_f$.

Estimation of the power consumed by friction on faults must consider both the thrust faulting at the overflow front and the extensional normal faulting of the continental interior. Friction on the thrust fault(s) at the front is likely to be low for two reasons: the hanging-wall side is subject to isostatic uplift forces (generated by the continental interior and transmitted through the 'fluid' ductile layer⁹) which reduce the fault normal stress; more importantly, hot volatiles at lithostatic pressure are likely to fill the fault¹⁰, typically derived from the overthrust wet sediments at the continental margin^{11–13}. These considerations suggest that the power consumed by thrusting at the overflow front can probably be neglected relative to the power consumed by the normal faults in the continental interior.

These normal faults do not need to have their power consumption evaluated individually; it is sufficient to consider only what tectonic force (compression reduction) must be applied at the leading edge of the continental interior to allow extension, and multiply that by the velocity to calculate the power consumption. The weight of experimental evidence^{14,15} suggests that maximum tensional stresses permitted by continental rocks before extension are well predicted by the wet Anderson stress deficit^{16,17}

$$\Delta \sigma_{xx}(z) = \xi \rho_c g z \quad (2)$$

where

$$\xi = -\frac{2k(1 - \rho_w/\rho_c)}{k + \sqrt{k^2 + 1}} \quad (3)$$

and k is the coefficient of friction for normal faulting. This stress is only supported by the brittle part of the lithosphere (in fact, this is the implicit definition of the brittle thickness); integration over the brittle thickness yields the force which may be multiplied by the frontal velocity to give, for the fault dissipation power per unit strike length, a value $P_n = (1/2)\xi \rho_c g b^2 v_f$. Because the spatial distribution of normal faults is irrelevant to this calculation, it applies not only to the case shown in Fig. 1, but also to the extreme case where normal faulting is confined to the continental border, so that lower crust is extruded from beneath a largely intact upper crust. For such extrusion, the lower-crustal flow regime would not be Couette (simple shear) flow as assumed above, but Poiseuille (channel) flow. The consequent eightfold increase in lower-crustal shear dissipation would not preclude this type of extrusion (see below), but would certainly make it less likely than the case shown in Fig. 1. The only other power sink linear in overflow velocity is the initial creation of flexural energy in bending the overridden oceanic lithosphere. I will neglect this also, as the total flexural energy, relative to available gravitational energy, must have been small for continents whose

width significantly exceeded the flexural length¹⁷ α of Archaean oceanic lithosphere. Even for modern cold oceanic lithosphere, α is of the order of 100 km, and would have been significantly less for hotter Archaean oceanic lithosphere.

Equating gravitational power input to fault dissipation yields a very simple criterion for the maximum continental elevation e_m above the ocean basins that can be stable against gravitational collapse: specifically, $e_m = \gamma b$ where $\gamma = (\xi \rho_c / \rho_g)^{1/2}$. For $\rho_c = 2,750 \text{ kg m}^{-3}$, $\rho_m = 3,300 \text{ kg m}^{-3}$, $\rho_w = 1,000 \text{ kg m}^{-3}$ and $k = 0.85$, γ is about one-third. At the time the typical elevation e slightly exceeded the mean ocean depth (assumed here to have been $\sim 5 \text{ km}$ even in Archaean times), the mean brittle thickness b of the continents must therefore have been of the order of 15 to 18 km. Assuming a brittle–ductile transition temperature of 400°C , the mean upper-crustal geothermal gradient must have dropped to somewhere between 22 and 28°C km^{-1} at the time gravitationally driven spreading ceased to be the norm. This value of gradient is currently achieved only in tectonically active regions of the Earth, but is not an unreasonable value for the norm at the close of the Archaean.

Although viscous shear dissipation does not determine the threshold for gravitational spreading, it can in principle render such a process so slow as to be tectonically irrelevant if the viscosities are too large. Mantle dissipation is not a problem in this sense, as indicated by the shortness of the timescale for glacial isostatic rebound ($\sim 10^4$ years). Crustal ductile-zone shear dissipation, for the simple shear flow consistent with the uniform thinning and extension model above, can be shown to be $(1/3)\eta_d L v_f^2/h$ where η_d and h are, respectively, the effective viscosity and thickness of the ductile zone. Estimates of the effective viscosity of the ductile channel in the deep crust using natural unloading of the deeper crust have typically been upper bounds. However, analysis¹⁸ of mining-related crustal unloading and uplift gave values of $6 \times 10^{17} \text{ Pa s}$ and 15 km , respectively, for the effective viscosity and thickness of the ductile channel, in a tectonically active region of southern Germany whose geothermal gradient (of the order of 30°C km^{-1})¹⁹ is comparable with what might have been the norm in Archaean times. With these values, the ductile shear dissipation is well under 1 per cent of the normal-fault dissipation, even for the tectonically significant overflow velocity of 3 cm yr^{-1} , and clearly does not impair the effectiveness of gravitational spreading in limiting continental elevation. In fact, the Archaean lower-crustal effective viscosity can be an order of magnitude larger than the above value (that is, $\sim 10^{19} \text{ Pa s}$) without invalidating the analysis presented here.

Gravitational spreading or overflow of continents onto ocean basins explains many distinctive features of the Archaean era. It would preclude stable passive margins in the modern sense. Overflow would juxtapose continental crust over oceanic crust (and marginal sediments) in the same way as shallow slab subduction, but without the subduction; shallow-subduction-style slab melting would then generate the tonalite–trondhjemite–granodiorite (TTG) rock suites²⁰ characteristic of Archaean terranes. The small but transient power requirements of initiation of thrusting and lithospheric flexing would probably make overflow episodic; the resulting fast intracontinental extensions produced by episodic overflow would also deplete by decompression melting the upper mantle immediately under the continental interior. Overflow tectonics would operate the lower crust as a giant gneiss factory, creating the background in which successively newer basaltic and TTG lithologies would be emplaced. The thermally determined cessation of overflow tectonics, and the concomitant stabilization of continental interiors against extension and of passive margins against overflow, defines the cratonization which marks the end of the Archaean era. Finally, this tectonic style is consistent with a recent proposal by Davis²¹, based on zircon ages, that the southern half of the Archaean Superior Province in Canada is overthrust on

metasediments, and not structurally the simple result of successive accretionary events. □

Received 17 November 1998; accepted 10 February 1999.

- de Wit, M. J. On Archean granites, greenstones, cratons and tectonics: does the evidence demand a verdict? *Precamb. Res.* **91**, 181–226 (1998).
- Hamilton, W. B. Archean magmatism and deformation were not products of plate tectonics. *Precamb. Res.* **91**, 143–179 (1998).
- McKenzie, D. & Fairhead, D. Estimates of the effective elastic thickness of the continental lithosphere from Bouger and free air gravity anomalies. *J. Geophys. Res.* **102**, 27523–27552 (1997).
- Dewey, J. F. Extensional collapse of orogens. *Tectonics* **7**, 1123–1139 (1988).
- Bird, P. Lateral extrusion of lower crust from under high topography. *J. Geophys. Res.* **96**, 10275–10286 (1991).
- Molnar, P. & Lyon-Caen, H. In *Processes in Continental Lithosphere Deformation* (eds Clark, S. P. Jr, Burchfiel, B. C. & Suppe, J.) 179–207 (Spec. Pap. 218, Geol. Soc. Am., Boulder, 1988).
- Artyushkov, E. V. J. Stresses in the lithosphere caused by crustal thickness inhomogeneities. *J. Geophys. Res.* **78**, 7675–7708 (1973).
- Turcotte, D. L. Mechanisms of crustal deformation. *J. Geol. Soc. Lond.* **140**, 701–724 (1983).
- Bailey, R. C. Thresholds for gravitationally induced thrusting adjacent to elevated topography. *Abstr. Geol. Soc. Am. Annu. Mtg* **30**, 297 (1998).
- Sibson, R. H. Frictional constraints on thrust, wrench and normal faults. *Nature* **249**, 542–544 (1974).
- Barr, T. D. & Dahlen, F. A. Constraints on friction and stress in the Taiwan fold-and-thrust belt from heat flow and geochronology. *Geology* **18**, 111–115 (1990).
- Bird, P. Initiation of intracontinental subduction in the Himalaya. *J. Geophys. Res.* **83**, 4975–4987 (1978).
- Bird, P. Stress and temperature in subduction zones: Tonga and Mariana. *Geophys. J. R. Astron. Soc.* **55**, 411–434 (1978).
- Zoback, M. D. & Harjes, H. P. Injection-induced earthquakes and crustal stress at 9 km depth at the KTB deep drilling site, Germany. *J. Geophys. Res.* **102**, 18477–18491 (1997).
- Brady, M., Zoback, M. D., Fuchs, K., Rummel, F. & Baumgartner, J. Estimation of the complete stress tensor to 8 km depth in the KTB scientific drill holes: implications for crustal strength. *J. Geophys. Res.* **102**, 18453–18475 (1997).
- Anderson, E. M. *The Dynamics of Faulting* (Oliver & Boyd, Edinburgh, 1951).
- Turcotte, D. & Schubert, G. *Geodynamics* (Wiley & Sons, New York, 1982).
- Klein, A., Jacoby, W. & Smilde, P. Mining induced crustal deformation in northwest Germany: modeling the rheological structure of the lithosphere. *Earth Planet. Sci. Lett.* **147**, 107–123 (1997).
- Hanel, R. in *Plateau Uplift* (eds Fuchs, K. et al.) 228–246 (Springer, Berlin, 1983).
- Martin, H. The mechanisms of petrogenesis of Archean continental crust comparison with modern processes. *Lithos* **30**, 373–388 (1993).
- Davis, D. W. Growth of the Superior craton: evidence from detrital zircon ages. *Abstr. Geol. Soc. Am. Annu. Mtg* **30**, 159 (1998).

Acknowledgements. I thank G. West for discussions of Archean tectonics. I also thank P. Bird for drawing my attention to the special case in which normal faulting is concentrated at the continental edge. This work was supported by the National Science and Engineering Research Council of Canada (NSERC).

Correspondence and requests for materials should be addressed to the author (e-mail: bailey@geophy.physics.utoronto.ca).

Absolute measures of the completeness of the fossil record

Mike Foote & J. John Sepkoski Jr

Department of the Geophysical Sciences, University of Chicago, 5734 South Ellis Avenue, Chicago, Illinois 60637, USA

Measuring the completeness of the fossil record is essential to understanding evolution over long timescales, particularly when comparing evolutionary patterns among biological groups with different preservational properties. Completeness measures have been presented for various groups based on gaps in the stratigraphic ranges of fossil taxa^{1,2} and on hypothetical lineages implied by estimated evolutionary trees^{3–5}. Here we present and compare quantitative, widely applicable absolute measures of completeness at two taxonomic levels for a broader sample of higher taxa of marine animals than has previously been available. We provide an estimate of the probability of genus preservation per stratigraphic interval^{6,7}, and determine the proportion of living families with some fossil record^{8–10}. The two completeness measures use very different data and calculations. The probability of genus preservation depends almost entirely on the Palaeozoic and Mesozoic records, whereas the proportion of living families with a fossil record is influenced largely by Cenozoic data. These measurements are nonetheless highly correlated, with outliers quite explicable, and we find that completeness is rather high for many animal groups.

The two measures of completeness lie along a positive trend except for the classes Cephalopoda and Chondrichthyes (Fig. 1). Most of the data fall below a line of unit slope; this is expected because genera consist of fewer species than do families, and because genus-level completeness is estimated per interval rather than for the entire genus duration. Two classes, Anthozoa and Malacostraca, lie slightly above this line. The low proportion of living anthozoan families with a fossil record largely reflects the many soft-bodied orders such as Actiniaria. The known fossil malacostracans, dominated by decapods, probably provide a small sample of all malacostracan groups. Estimated preservation probability for brachiopod genera slightly exceeds the theoretical maximum of unity in one analysis, but the excess is well within sampling error based on resampling methods or on the multinomial distribution⁷. The proportion of living polychaete families with a fossil record is moderately high, but many of these are tube-secreting worms¹¹ that have diversified toward the Recent; polychaete genera from the Mesozoic and especially the Palaeozoic include many fossils known from single localities.

The explanation as to why two classes deviate from the positive trend in Fig. 1 is provided by the nature of their fossil record and the composition of the living fauna (Fig. 2). Regarding the Cephalopoda, except for the few surviving nautilid species, living cephalopods belong to the subclass Coleoidea, most of which are soft or lightly skeletonized; this accounts for the low proportion of living families with a fossil record. Fossil taxa are primarily from well-skeletonized groups, namely the once diverse nautiloids and the extinct ammonoids. When we divide the fossil genera into three subclasses, estimated preservation probability is lowest for coleoids (Fig. 2a). The high range of coleoid values partly reflects sampling variance⁷,

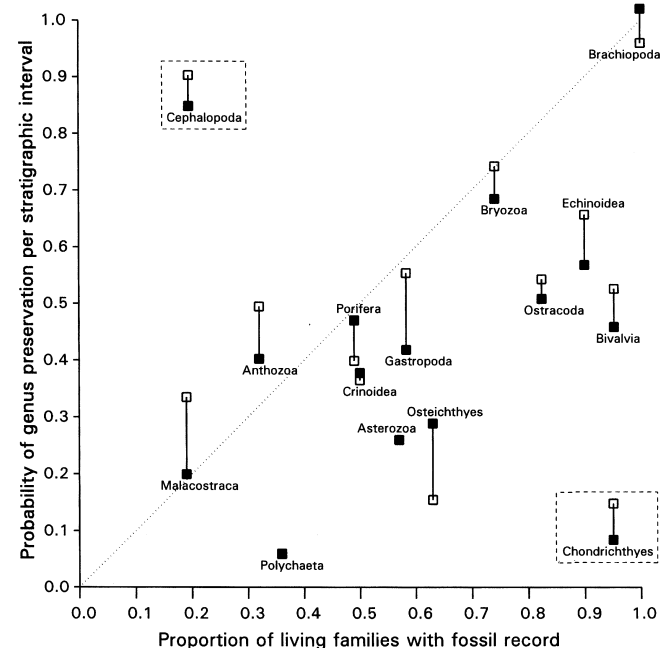


Figure 1 Absolute measures of completeness for higher taxa of marine animals. For each taxon, the proportion of living families with a fossil record is compared with the estimated probability of genus preservation per stratigraphic interval. The data show a positive trend, indicating a correlation between the different completeness measures at two taxonomic levels and temporal scopes of data (genera mainly pre-Cenozoic; families mainly Cenozoic). Taxa that deviate from the trend are enclosed in boxes and analysed in Fig. 2. Two measures of genus-level completeness, reflecting different subdivisions of the Phanerozoic, are shown connected by a bar. The open squares denote the 103-interval timescale, whereas the closed squares denote the 107-interval timescale. The dotted diagonal line has unit slope.

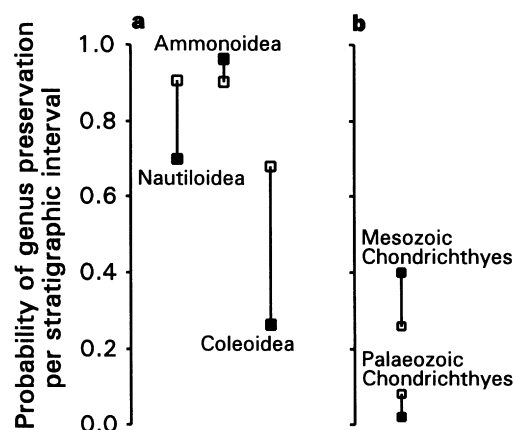


Figure 2 Genus-level completeness for subgroups within higher taxa that deviate from the trend presented in Fig. 1. **a**, For cephalopods, Coleoidea (containing nearly all living forms, most of which are soft-bodied) has the lowest completeness, whereas completeness of the well-skeletonized subclasses (the extinct Ammonoidea and the now depauperate Nautiloidea) is higher. **b**, Among chondrichthyans, Palaeozoic forms are known mainly from whole-body fossils confined to exceptional deposits, whereas Mesozoic forms are more similar to their modern counterparts and have a well-documented fossil record of teeth. Overall genus completeness is dominated by the single-interval Palaeozoic forms (75% of Palaeozoic genera are confined to single intervals, compared with 52% for Mesozoic genera). Key to symbols as in Fig. 1.

as there are only about 120 fossil genera in our calculations, and it partly reflects the heterogeneity⁷ in the data, which arises from including the well-preserved belemnites. The higher completeness of ammonoids relative to nautiloids probably reflects their biostratigraphic utility as zonal index fossils and consequently their intensive study.

The proportion of living chondrichthyan families with a fossil record is high because many families can be extended back in time on the basis of fossil teeth. It is instructive to divide the fossil genera into those first appearing during the Palaeozoic Era and those first appearing during the Mesozoic (Fig. 2b). Many Palaeozoic forms are whole-body fossils from deposits with exceptional preservation¹². These are thus single-interval taxa, contributing to low estimates of completeness (see Methods). Mesozoic forms, like the fossil representatives of their living counterparts, tend more to be described from teeth outside of exceptional fossil deposits¹³; they thus have longer ranges and higher completeness values.

In addition, we calculated genus-level completeness for some palaeontologically important, extinct higher taxa (Fig. 3). Blastozoan echinoderms have a fossil record more complete than that of asterozoans (many of which disarticulate shortly after death), comparable to that of crinoids (which share several ecological and structural features with blastozoans), and less complete than that of echinoids (which include the robust irregulars that diversified greatly during the Mesozoic). The biostratigraphically important trilobites, graptolites and conodonts have rather complete fossil records at the genus level. This result reinforces the hypothesis that there is a positive feedback between biostratigraphic utility and completeness. In addition to evolving quickly and being geographically widespread, biostratigraphically useful taxa should have high intrinsic preservability. Such taxa are also well documented in global databases.

Completeness reflects taxonomic duration as well as intrinsic preservability. With all other factors equal, taxa with shorter total durations are less likely to enter the fossil record, and taxa that endure through a smaller proportion of a time interval have a lower probability of preservation during that interval^{2,6,7}. There are nonetheless several clades that have high completeness despite short

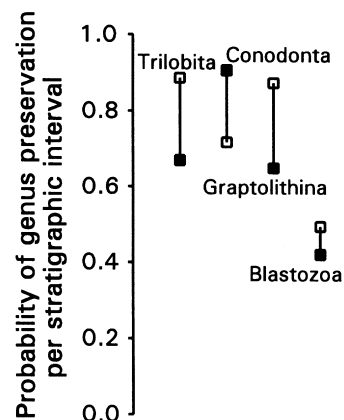


Figure 3 Genus-level completeness for major, extinct higher taxa. The biostratigraphically important trilobites, conodonts and graptolites have high completeness. The completeness of blastozoan echinoderms is higher than that of asterozoan echinoderms, comparable to that of crinoids, and lower than that of echinoids (Fig. 1). Key to symbols as in Fig. 1.

durations (Trilobita, Graptolithina, Conodonta and Ammonoidea) as well as groups with low to intermediate completeness despite long durations (Malacostraca, Porifera and Mesozoic Chondrichthyes). This further indicates a distinct signal of preservability in our two sets of data.

Many of the genera and families in this study are paraphyletic. Such taxa are often regarded as artefacts that add noise¹⁴, blurring a true palaeontological signal. We can think of no reason that noise, whether it reflects taxonomic practice or stratigraphic error, should create, rather than blur, the trend we see in our two measures of completeness. That we can even detect such a trend despite the imperfections in our data indicates that the trend is real. Our methods do have limitations^{1,2,6,7}, such as the tacit assumption of homogeneous preservation within higher taxa. However, we know of no bias that would cause two very different methods applied to two different data sets to show the same pattern. We therefore conclude that the two methods are detecting a common signal of preservability and that taxa at different hierarchical levels in the traditional Linnaean taxonomy are capturing similar information from the fossil record.

Our method of estimating the probability of genus preservation has been used to determine whether apparent differences in evolutionary rates are likely to be artefacts of differential completeness⁶. Incompleteness affects other palaeontological problems, including sudden versus gradual extinction patterns^{15,16}, evolutionary relationships^{17–21}, the temporal order of appearance of taxa^{1,21–23} and taxonomic diversity⁹. In tabulations of fossil diversity, taxa with very different degrees of preservation (for example, brachiopods and polychaetes) have simply been summed. Our quantitative measures suggest how different groups might be weighted in global diversity studies. Thus, these measures and others^{2–5,15} may aid an informed reading of evolution in the fossil record. □

Methods

Genus-level preservation probability. Our estimate of preservation probability per stratigraphic interval is derived from the following result. If f_1 , f_2 and f_3 are the frequencies of genera with stratigraphic ranges of one, two and three intervals of geologic time, then the ratio $f_2^2/(f_1 f_3)$ estimates the per-interval preservation probability^{6,7}. The lower the preservation probability, the more taxa are confined to single time intervals, and thus the lower is this ratio. In common with many approaches to estimating preservation probability^{1,2}, this measure tacitly assumes homogeneous data, a condition we attempt to satisfy by analysing major taxa independently. Violations of this assumption and various statistical properties of this completeness measure are discussed elsewhere^{2,6,7}.

Timescales. To measure genus-level completeness we used two alternative stratigraphic subdivisions of the Phanerozoic, developed independently by the two authors. The first was a previously published subdivision²⁴ using various timescales and conventional stratigraphic series, stages and substages split and lumped to yield 103 units with a mean duration of 5.5 million years (Myr) and approaching minimal variance in duration. The second was a subdivision based mainly on the timescale of ref. 25, variously divided to yield 107 intervals averaging 5.3 Myr in duration, again approaching minimal variance. Although the 107-interval subdivision tends to yield lower estimates of preservation probability (Figs 1–3), the tendency is not statistically significant by a binomial test ($N = 17$ cases in which the two timescales yield different results; two-tailed $P = 0.14$). The dependence of estimated preservation probability on timescale reflects idiosyncrasies in the lumping and splitting of stratigraphic units rather than the average level of resolution, which differs little between the two timescales. With each subdivision, we included all genera in a comprehensive, global database^{26,27} whose first and last occurrences could be reliably placed into one of the stratigraphic intervals. Out of 24,190 genera in the taxonomic groups analysed, 16,464 (68.1%) could be resolved with the 103 intervals of geologic time, and 16,338 (67.5%) could be resolved with the 107 intervals.

Data. We used the family data compiled by one of us^{26,28}, supplemented extensively with data from ref. 29, to measure the proportion of living families with some fossil record. The taxonomy of living families and the number of unfossilized living families are based largely on ref. 30. To limit problems of the 'Pull of the Recent'⁹ in the analysis of genera, including the artificial truncation of durations that would affect our estimate of genus preservation probability^{6,7}, we used only genera that first occur before the Cenozoic era. We excluded groups with less than 100 resolvable fossil genera or less than 10 living families, which, with greater proportional counting error, would yield less stable measures of completeness. Our data thus favour more diverse and better-preserved groups (nematodes, priapulids and many other soft-bodied phyla were excluded, for example), but our measures are not biased for those groups used as we included all resolvable marine taxa within each group. We did not tabulate the proportion of living genera with a fossil record, as adequate data are not available for most groups.

Sample sizes for Fig. 1 (number of living families and number of genera resolved with the 103-interval timescale and the 107-interval scale) are: Anthozoa (122, 1,179 and 1,162); Asterozoa (53, 154 and 156); Bivalvia (103, 833 and 804); Brachiopoda (19, 2,840 and 2,778); Bryozoa (143, 715 and 734); Cephalopoda (41, 3,133 and 3,114); Chondrichthyes (42, 173 and 169); Crinoidea (26, 824 and 830); Echinoidea (50, 228 and 241); Gastropoda (230, 885 and 873); Malacostraca (371, 191 and 201); Osteichthyes (328, 361 and 312); Ostracoda (34, 1,072 and 1,073); Polychaeta (75, 123 and 123); and Porifera (107, 590 and 570). Because no asterozoan genera have resolved ranges of three intervals on the 107-interval scale, results are presented for the 103-interval scale only. Porifera excludes Archaeocyatha, which are very short-lived as a group and thus have durations severely truncated by the extinction of the group shortly after its origin. Sample sizes for Figs 2 and 3 (number of genera in 103-interval scale and in 107-interval scale) are: Ammonoidea (2,235 and 2,215); Blastozoa (237 and 237); Coleoidea (127 and 118); Conodonts (261 and 264); Graptolithina (226 and 229); Nautiloidea (781 and 791); Trilobita (2,439 and 2,468); Palaeozoic Chondrichthyes (82 and 83); and Mesozoic Chondrichthyes (91 and 86). To increase sample size for coleoids, genera originating in the Palaeocene and Eocene have been included.

Received 17 August 1998; accepted 8 February 1999.

1. Paul, C. R. C. in *Problems of Phylogenetic Reconstruction* (eds Joysey, K. A. & Friday, A. E.) 75–117 (Academic, London, 1982).
2. Solow, A. R. & Smith, W. On fossil preservation and the stratigraphic ranges of taxa. *Paleobiology* 23, 271–277 (1997).
3. Benton, M. J. & Storrs, G. W. Testing the quality of the fossil record: paleontological knowledge is improving. *Geology* 22, 111–114 (1994).
4. Benton, M. J. & Simms, M. J. Testing the marine and continental fossil records. *Geology* 23, 601–604 (1995).
5. Benton, M. J. & Hitchin, R. Testing the quality of the fossil record by groups and by major habitats. *Histor. Biol.* 12, 111–157 (1996).
6. Foote, M. & Raup, D. M. Fossil preservation and the stratigraphic ranges of taxa. *Paleobiology* 22, 121–140 (1996).
7. Foote, M. Estimating taxonomic durations and preservation probability. *Paleobiology* 23, 278–300 (1997).
8. Schopf, T. J. M. Fossilization potential of an intertidal fauna: Friday Harbor, Washington. *Paleobiology* 4, 261–270 (1978).
9. Raup, D. M. Biases in the fossil record of species and genera. *Bull. Carnegie Mus. Nat. Hist.* 13, 85–91 (1979).

10. Valentine, J. W. How good was the fossil record? Clues from the Californian Pleistocene. *Paleobiology* 15, 83–94 (1989).
11. Wills, M. A. Annelids, in *The Fossil Record* Vol. 2 (ed. Benton, M. J.) 271–278 (Chapman & Hall, London, 1993).
12. Zangerl, R. in *Handbook of Paleichthyology* Vol. 3A (ed. Schultze, H.-P.) 1–115 (Gustav Fischer, Stuttgart, 1981).
13. Cappetta, H. in *Handbook of Paleichthyology* Vol. 3B (ed. Schultze, H.-P.) 1–193 (Gustav Fischer, Stuttgart, 1987).
14. Patterson, C. & Smith, A. B. Is the periodicity of extinction a taxonomic artefact? *Nature* 330, 248–251 (1987).
15. Marshall, C. R. & Ward, P. D. Sudden and gradual molluscan extinctions in the latest Cretaceous of western European Tethys. *Science* 274, 1360–1363 (1996).
16. Signor, P. W. & Lipps, J. H. Sampling bias, gradual extinction patterns, and catastrophes in the fossil record. *Geol. Soc. Amer. Spec. Pap.* 190, 291–296 (1982).
17. Fortey, R. A. & Jefferies, R. P. S. in *Problems of Phylogenetic Reconstruction* (eds Joysey, K. A. & Friday, A. E.) 197–234 (Academic, London, 1982).
18. Wagner, P. J. Stratigraphic tests of cladistic hypotheses. *Paleobiology* 21, 153–178 (1995).
19. Fisher, D. C. in *Interpreting the Hierarchy of Nature: From Systematic Patterns to Evolutionary Processes* (eds Grande, L. & Rieppel, O.) 133–171 (Academic, San Diego, 1994).
20. Foote, M. On the probability of ancestors in the fossil record. *Paleobiology* 22, 141–151 (1996).
21. Paul, C. R. C. The adequacy of the fossil record reconsidered. *Spec. Pap. Paleontol.* 33, 7–16 (1985).
22. Norell, M. A. & Novacek, M. J. The fossil record and evolution: comparing cladistic and paleontological evidence for vertebrate history. *Science* 255, 1690–1693 (1992).
23. Hitchin, R. & Benton, M. J. Congruence between parsimony and stratigraphy: comparisons of three indices. *Paleobiology* 23, 20–32 (1997).
24. Sepkoski, J. J. Jr in *Global Events and Event Stratigraphy* (ed. Walliser, O.) 35–52 (Springer, Berlin, 1996).
25. Harland, W. B. et al. *A Geologic Time Scale 1989* (Cambridge Univ. Press, 1990).
26. Sepkoski, J. J. Jr Biodiversity: past, present, and future. *J. Paleontol.* 71, 533–539 (1997).
27. Sepkoski, J. J. Jr Rates of speciation in the fossil record. *Phil. Trans. R. Soc. Lond. B* 353, 315–326 (1998).
28. Sepkoski, J. J. Jr A compendium of fossil marine animal families, 2nd edition. *Milwaukee Publ. Mus. Contrib. Biol. Geol.* 83, 1–156 (1992).
29. Benton, M. J. (ed.) *The Fossil Record* Vol. 2 (Chapman & Hall, London, 1993).
30. Parker, S. P. (ed.) *Synopsis and Classification of Living Organisms* (McGraw-Hill, New York, 1982).

Acknowledgements. We thank E. G. Hunt, A. McGowan, A. I. Miller, and P. J. Wagner for comments. This work was supported by the US National Aeronautics and Space Administration and the US National Science Foundation.

Correspondence and requests for materials should be addressed to M.F. (e-mail: mfoote@midway.uchicago.edu).

Synthesis of thrombin-inhibiting heparin mimetics without side effects

Maurice Petitou, Jean-Pascal Hérault, André Bernat, Pierre-Alexandre Driguez, Philippe Duchaussoy, Jean-Claude Lormeau & Jean-Marc Herbert*

Sanofi Recherche, 195 route d'Espagne, 31036 Toulouse, France

Unwanted side effects of pharmacologically active compounds can usually be eliminated by structural modifications. But the complex heterogeneous structure of the polysaccharide heparin¹ has limited this approach to fragmentation, leading to slightly better-tolerated heparin preparations of low molecular mass². Despite this improvement, heparin-induced thrombocytopenia³ (HIT), related to an interaction with platelet factor 4 (PF4) and, to a lesser extent, haemorrhages⁴, remain significant side effects of heparinotherapy. Breakthroughs in oligosaccharide chemistry⁵ made possible the total synthesis of the pentasaccharide anti-thrombin-binding site of heparin^{6,7}. This pentasaccharide represents a new family of potential antithrombotic drugs, devoid of thrombin inhibitory properties, and free of undesired interactions with blood and vessel components. To obtain more potent and well-tolerated antithrombotic drugs, we wished to synthesize heparin mimetics able to inhibit thrombin, that is, longer oligosaccharides. Like thrombin inhibition, undesired interactions are directly correlated to the charge and the size of the molecules⁸, so we had to design structures that were able to discriminate between thrombin and other proteins, particularly PF4. Here we describe the use of multistep converging synthesis to obtain sulphated oligosaccharides that meet these requirements.

The structures we synthesized were inspired by the structure of heparin itself and designed according to current knowledge of its

mechanism of action⁹ (Fig. 1). Briefly, heparin binding to antithrombin induces a conformational change in the inhibitory loop of the protein (allosteric activation) resulting in recognition and inhibition of factor Xa. Thrombin is inhibited through a template mechanism (entropic activation): it is electrostatically attracted by heparin and collides with heparin-bound antithrombin. Thus, thrombin inhibition requires a longer heparin chain, more prone to interact with other biological molecules, particularly PF4, than the short pentasaccharide⁸. Natural heparin chains, because of their sulphation profile, cannot discriminate between thrombin and many other proteins (hence the undesired side effects). But chemical synthesis, through the fine tuning of the length and the charge of the synthetic oligosaccharides, should allow us to identify structures able to do so. Our approach was also influenced by our ultimate goal, to obtain a drug substance, which pushed us to keep the chemistry as simple as possible to facilitate scaling up.

The above mechanistic considerations on the anticoagulant action of heparin suggest that the structure of an oligosaccharide molecule able to mimic its full anticoagulant activity would include an antithrombin-binding domain (A-domain) coupled to a thrombin-binding domain¹⁰ (T-domain). Such a heparin sequence comprises between 14 and 20 saccharide units^{11–14}. Moreover, from the specific interaction of antithrombin with heparin (Fig. 1c, d), it was inferred that one only of the two possible ways of elongating the A-domain with a T-domain (either at the reducing or at the non-reducing end) should lead to thrombin inhibition¹⁰. Our first approach towards such molecules, dictated by our desire to use a 'short' synthetic route, circumvented the problem of the relative position of these two domains. We reasoned that an A-domain, because it is negatively charged, can serve as a T-domain, and that a continuum of A-domains would necessarily display the proper relative position of the two domains, because antithrombin could bind at either end of the molecule, and thrombin would be attracted by the remaining part of the chain (Fig. 1). Thrombin inhibition should therefore be observed as soon as the saccharide is long enough to accommodate antithrombin and thrombin at the same time. We worked with the so-called 'non-glycosaminoglycan'

series¹⁵, in which *N*-sulphate groups are replaced by *O*-sulphate groups, and hydroxyls are alkylated (Fig. 2). These structural modifications fully preserve the specific binding to antithrombin¹⁶ yet they dramatically simplify the synthesis.

Having identified a highly symmetrical hexasaccharide that can bind to and activate antithrombin, we synthesized larger homologous oligosaccharides and tested their anticoagulant properties¹⁷. All compounds, because they contain an A-domain, bound to antithrombin and inhibited factor Xa. The 6-, 10-, 12- and 14-mer saccharides were inactive in the thrombin inhibition assay, whereas the 16-, 18- and 20-mer saccharides displayed size-dependent increasing activity in this assay, the 20-mer saccharide being half as potent as standard heparin (Table 1). In an *in vivo* experimental model of venous thrombosis (Wessler model in the rat)¹⁸, in which thrombosis was induced by a combination of a thrombogenic challenge (recombinant tissue factor) and stasis, all compounds displayed a similar, dose-dependent antithrombotic effect related to their specific anti-factor Xa activity (Table 1). In an arterial thrombosis model (arterio-venous shunt in the rat) in which platelets and thrombin have a preponderant role¹⁹, the intravenous administration of the thrombin inhibitory compounds (5–7) reduced thrombus formation in a dose-dependent manner with half-maximal effective dose (ED₅₀) values similar to that of standard heparin, whereas compounds 1–6 (devoid of thrombin inhibitory effect) were inactive (Table 1). Unfortunately, all these molecules were neutralized by PF4 and they activated platelets in the presence of plasma from HIT patients (not shown). Thus, despite their antithrombotic properties (the 20-mer saccharide exhibited an antithrombotic activity similar to that obtained for standard heparin), the cross reactivity of these compounds with PF4 prompted us to explore another family of related molecules with a structure closer to that of heparin, that is, possessing a specific A-domain prolonged by a T-domain that is not recognized by antithrombin (Fig. 2).

The above results on thrombin inhibition by compounds 2–7 clearly established that the oligosaccharide we are looking for must contain at least 15 or 16 saccharide units (this gives a more precise answer to the long-standing question about the size of heparin fragments able to catalyse thrombin inhibition by antithrombin^{11–14}). In contrast to the first approach, an important issue in the design of these structures was how to attach the T-domain at the proper end of the A-domain to obtain efficient thrombin inhibition (Fig. 1). Studies based on modelling of the ternary heparin/antithrombin/thrombin complex¹⁰ and recent crystallography studies²⁰ suggested that the T-domain may need to be attached at the non-reducing end so, knowing that we had to synthesize at least a 15-mer saccharide, we targeted compounds 8–10. We selected as A-domain a high-affinity analogue of the antithrombin-binding sequence²¹ (DEFGH; Fig. 2). Because thrombin binding in the T-domain is mainly a matter of electrostatic attraction of the anion-binding exosite II of the protein by the anionic polysaccharide²², and in order to keep the chemistry feasible, we used alternating α - and β -linked 3-*O*-methyl-2,6-di-*O*-sulpho-D-glucose units as a mimic of the regular region of heparin. Biological tests performed on these compounds (Table 1) demonstrated that they were inhibitors of both factor Xa and thrombin. Thrombin inhibition was size dependent, which explains the greater ability of a longer negatively charged molecule of heparin to attract thrombin and bring it into contact with antithrombin. The minimum number of saccharide units in the chain that could inhibit thrombin was 15 (compound 8). Note that the 19-mer saccharide (compound 10) was as potent as the most active fraction isolated from a standard heparin preparation²³.

We then had to prove that the opposite arrangement of A- and T-domains resulted in an inactive compound. We therefore synthesized the 18-mer saccharide 11 (Fig. 2) made of the same structural elements as 8–10 but in which their relative positions have been

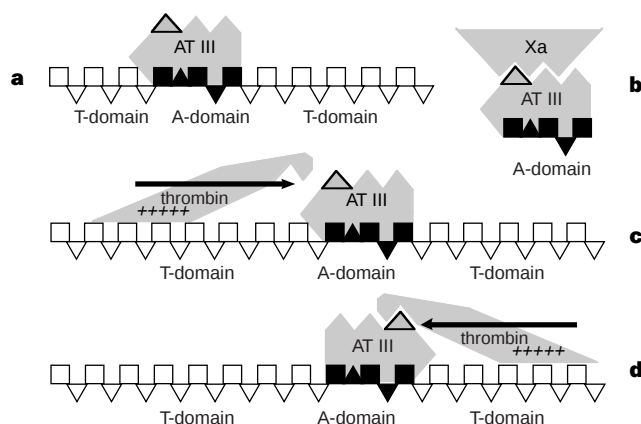


Figure 1 Inhibition of factor Xa and thrombin by antithrombin and heparin. Heparin is shown here (see also Fig. 2) as a repetition of squares (glucosamine units) and triangles (uronic acid units). **a**, Through binding to the antithrombin pentasaccharide-binding-site (A-domain, black squares and triangles), antithrombin conformation changes (hatched triangle) in a manner sufficient to inhibit factor Xa without any contribution from the rest of the saccharidic chain. This is why the pentasaccharide representing the antithrombin-binding site can inhibit factor Xa (**b**). In contrast, thrombin inhibition requires attraction by the negative charges of the thrombin-binding domain (T-domain). Thrombin 'slides' along the heparin chain until it hooks itself on the exposed loop of activated antithrombin. Because the heparin-antithrombin complex has a well-defined structure, thrombin must approach from the correct side (possibly **c** or **d**) to be irreversibly inhibited.

Table 1 Biological properties of the synthetic oligosaccharides

Compound number	1	2	3	4	5	6	7	8	9	10	11	12	Heparin	LMWH
Number of saccharide units	6	10	12	14	16	18	20	15	17	19	18	17	~10–50	~6–30
Molecular mass	2,217	3,606	4,301	4,995	5,690	6,384	7,078	5,618	6,378	7,139	6,698	5,322	~15,000	~4,500
<i>In vitro</i> activities														
Affinity for antithrombin (K_d , nM)	350 ±10	21 ±4.3	19 ±2.9	19 ±2.1	22 ±1.3	28 ±1.4	24 ±1.7	16 ±0.3	3.3 ±0.8	1.2 ±0.2	7.3 ±1.4	7.0 ±1.5	25 ±0.2	n.d.*
Factor Xa inhibition (units mg^{-1})	325 ±16	405 ±32	360 ±29	310 ±16	359 ±29	270 ±23	236 ±19	370 ±9	270 ±8	290 ±29	230 ±16	270 ±8	180 ±0.5	170 ±17
Thrombin inhibition (IC_{50} , ng ml^{-1})	>10,000	>10,000	>10,000	>10,000	130 ±10	23 ±4	6.7 ±3	41 ±3	5.3 ±0.2	1.7 ±0.5	164 ±6.5	5.3 ±0.3	3.3 ±0.5	27 ±
<i>In vivo</i> activities														
Venous thrombosis (ED_{50} , $\mu\text{g kg}^{-1}$, i.v.)	300 ±70	280 ±49	360 ±62	390 ±45	150 ±21	250 ±20	110 ±17	65.5 ±3	40 ±4	38 ±9	423 ±27	15 ±2	80 ±3	87 ±3
Arterial thrombosis (ED_{50} , $\mu\text{g kg}^{-1}$, i.v.)	>1,000	>1,000	>1,000	>1,000	520 ±140	610 ±230	620 ±130	380 ±30	570 ±130	770 ±200	>1,000	70 ±16	700 ±90	>1,000

Means $\pm$ s.d. ($n = 10$). Affinity for antithrombin, factor Xa inhibition, thrombin, inhibition and anti-thrombotic activities were determined using published procedures¹⁸. MWH, low-molecular-weight heparin.

* n.d., not determined.

reversed. The biological properties of compound **11** (Table 1) showed that the affinity for antithrombin and the anti-factor Xa activity were in the same range as that observed for compounds **8–10**. However, compound **11** hardly inhibited thrombin in the presence of antithrombin (from 30 to 100 times weaker on a mass basis than the corresponding 17/19-mer saccharides **9** and **10**).

As seen with compounds **5–7**, although compounds **8–10** were

potent antithrombin-mediated thrombin inhibitors, their anti-coagulant activity was neutralized by PF4, indicating that reducing the size to the minimum that still allowed thrombin inhibition was not sufficient to abolish the undesired interaction with PF4.

We therefore explored how to reduce the charge of the molecule, the second parameter governing nonspecific reactivity of polyanionic compounds⁸. We based the design of our new targets on the

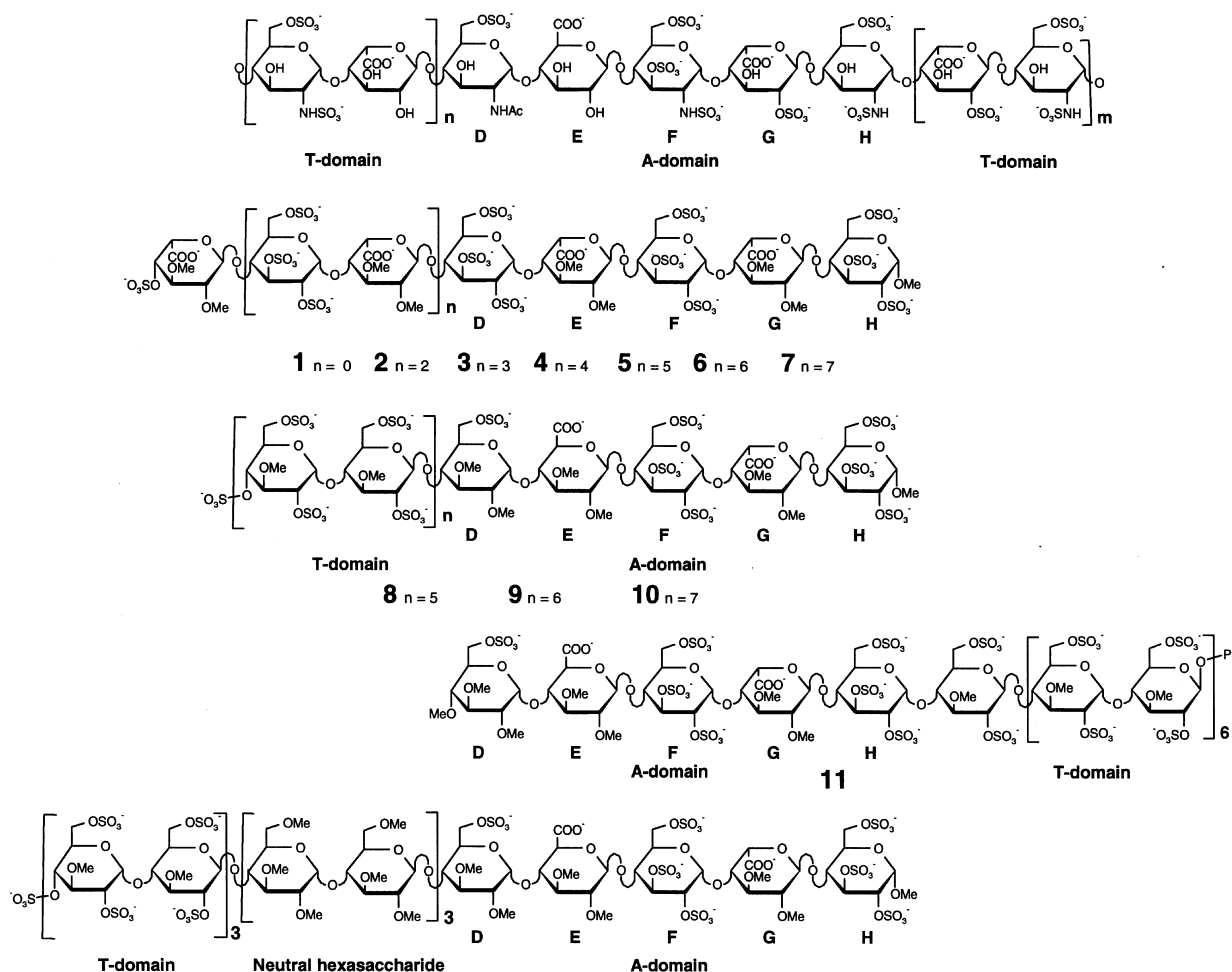


Figure 2 Structure of an anticoagulant heparin molecule (top line) and of oligosaccharides 1–12.

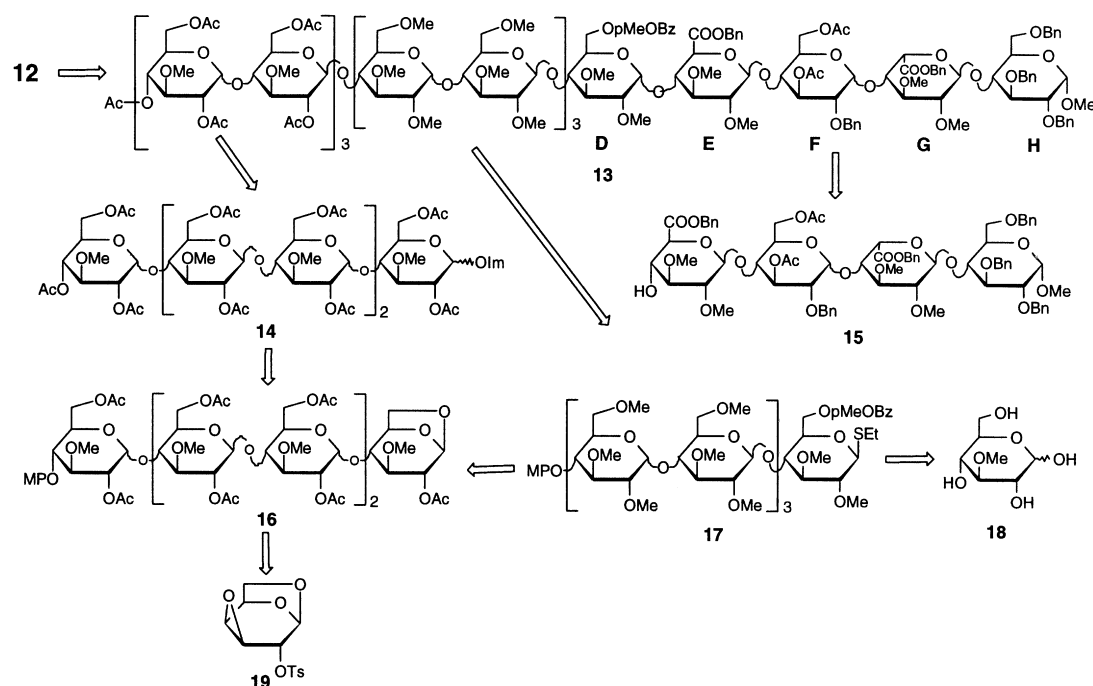


Figure 3 Retrosynthetic analysis of compound 12.

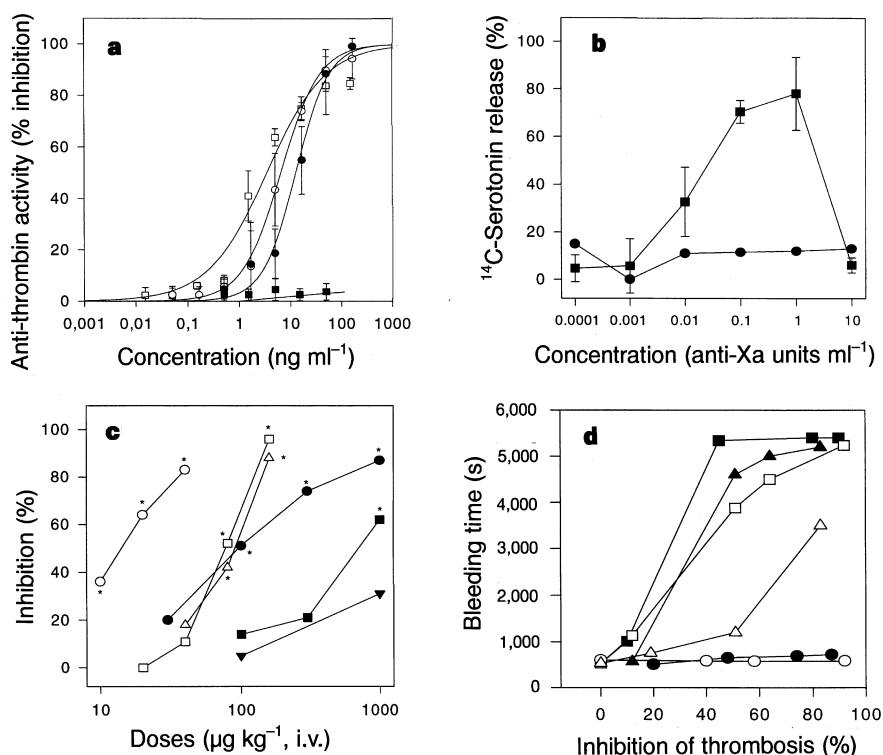


Figure 4 Biological properties of compound 12. **a**, Inhibitory activity for thrombin. Human thrombin was incubated with antithrombin in the presence of compound 12 (circles) or standard heparin (squares) either with (filled symbols) or without (unfilled symbols) PF4. The percentage of inhibition of thrombin activity was calculated from three experiments performed in triplicate. **b**, Effect of compound 12 on platelet activation in the presence of plasma from HIT patients. Increasing concentrations of compound 12 (circles) or standard heparin (squares) were incubated with washed human platelets previously incubated with ¹⁴C-serotonin in the presence of plasma from HIT patients. Platelet activation was measured by ¹⁴C-serotonin release. Each point is the mean $\pm$ s.d. ($n = 6$). **c**, Venous and arterial thrombosis in the rat. The antithrombotic activity of compound 12 (circles),

standard heparin (squares) or low-molecular-mass heparin (triangles) was determined in venous (unfilled symbols) and arterial (filled symbols) models of thrombosis. Each point is the mean of 10 animals per group. Grouped data were analysed for significance by comparison with the vehicle-treated group using the Mann-Whitney test with Holm-Bonferroni adjustment; $*P < 0.05$. **d**, Bleeding time in the rat. Increasing doses of compound 12 (circles), standard heparin (squares) or low-molecular-mass heparin (triangles) were administered i.v. to rats. Results are expressed as mean bleeding times as a function of the antithrombotic activity of the considered compound in a venous (unfilled symbols) or arterial thrombosis model (filled symbols). Each point is the mean of 10 animals per group.

following considerations: (1) a pentasaccharide sequence (an A-domain) is required to bind and activate antithrombin towards factor Xa and thrombin inhibition; (2) the T-domain must be two to three disaccharides long²⁴; (3) a chain length of 17 saccharide units is required for notable thrombin inhibition; and (4) the six or eight units of the central saccharide in compound **9** are not critically involved in the interaction either with antithrombin or with thrombin, and the charges on these units might thus be suppressed without affecting the anticoagulant activity. This analysis, which was supported by modelling studies of the ternary antithrombin/thrombin/heparin complex¹⁰, led us to synthesize (Fig. 3) compound **12** (Fig. 2), which comprises an A-domain having a very high affinity for antithrombin²¹ and a sulphated hexasaccharide as a T-domain, separated by a neutral methylated hexasaccharide.

It has been reported that a heparin fragment must be an 8-mer saccharide or a larger fragment²⁵ to get significant interaction with PF4. But the two charged domains in our structure are below this critical length. As might be expected from this observation, the thrombin inhibitory potency of compound **12** (Table 1, Fig. 4) was not affected by PF4, even when added at a very high concentration (100 µg ml⁻¹; Fig. 4a). This may be the reason why compound **12** did not activate platelets in the presence of plasma from HIT patients (Fig. 4b), suggesting that it will not induce thrombocytopenia. This may similarly explain why we observed an increased antithrombotic activity compared to heparin and the other synthetic derivatives described here. Thus, in models of both venous and arterial thrombosis, compound **12** was 5- to 10-fold more potent than standard heparin and low-molecular-mass heparin (Table 1, Fig. 4c). Another aspect of such a strong decrease of nonspecific interactions with basic proteins (such as von Willibrand factor and fibrinogen, the interaction that may participate in the haemorrhagic activity of heparin) is the lack of haemorrhagic activity of compound **12** compared with standard and low-molecular-mass heparin (Fig. 4d). The reduction in bleeding time observed in this experimental model may well augur a lower haemorrhagic effect observed in patients with this compound, compared with standard and low-molecular-mass heparin. Finally, from a pharmacokinetics standpoint, the compounds described here had a very simple elimination profile compared with that of standard heparin and low-molecular-mass heparin. This is probably related to the limited number of biological molecules with which they can interact, but also the structural homogeneity of these compounds obtained by chemical synthesis.

Taken together, our results demonstrate that substitutes for heparin, endowed with potent antithrombotic activity but devoid of its major side effects, can be obtained by chemical synthesis. Clinical trials will tell whether such synthetic mimetics can replace heparin, which is more than half a century old. □

Methods

Oligosaccharide synthesis. The general strategy for the synthesis of the sulphated oligosaccharides **1–12** relied on elaboration of the appropriate carbohydrate backbone adequately substituted by protective groups. After removal of the protective groups, the free hydroxyl groups were sulphonated using triethylamine-sulphur trioxide complexes. High-field ¹H-nuclear magnetic resonance (500 MHz), electrospray ionization mass spectrometry, and capillary electrophoresis were used to check the structure and the homogeneity of the compounds that were 90–97% pure. Compounds **1–7** were efficiently obtained from a single disaccharide building-block¹⁷. A retrosynthetic analysis of the synthesis of compound **12** is shown in Fig. 3. The fully protected oligosaccharide **13** is a synthetic equivalent of **12**. It was obtained from tetrasaccharide **15**, which was glycosylated at position H-4'' by reaction with the thioglycoside glycosyl donor **17**. After selective removal of the *p*-methoxyphenyl protecting group, a new glycosylation reaction using the imidate **14** was carried out, which delivered **13**. To minimize the number of steps, **14** and **15** derived from the same hexasaccharide precursor **16**, the key synthon of this synthesis. A *p*-methoxybenzoyl ester was chosen to protect position 6 of the D

unit in **17** because we observed that such an ester strongly orientated a glycosylation reaction towards the formation of the α-anomer, when a benzyl ether was present at position 2 of glucose. The choice of a *p*-methoxyphenyl ether to protect position four of the non-reducing end unit of **16** was dictated by the many different types of reaction conditions the temporary protective group at this position had to withstand. The synthesis of **8–11** used a similar route but did not require the methylated synthon **17**.

Inhibitory activity for thrombin. Purified human α-thrombin (2 U ml⁻¹) was incubated for 1 min with human antithrombin (0.25 U ml⁻¹) at 37 °C in the presence of the various compounds with or without 10 µg ml⁻¹ human PF4. To measure the residual thrombin activity, S-2238 chromogenic substrate was added (100 µM), the reaction was stopped after 1 min and the absorbance at 405 nm (A₄₀₅) was read. The percentage of inhibition of thrombin activity was calculated from three experiments performed in triplicate.

Inhibitory activity for factor Xa. Anti-factor Xa activities were determined by a modification of the Teien and Lie procedure as described¹⁸. Briefly, human factor Xa (2.4 nkat ml⁻¹) was incubated for 2 min with human antithrombin (0.17 U ml⁻¹) at 37 °C in the presence of various concentrations of the oligosaccharides in Tris/maleate 20 mM buffer pH 7.4, NaCl 150 mM. To measure the residual factor Xa activity, S-2222 (dissolved in 50 mM Tris-HCl buffer, pH 8.4, NaCl 175 mM, EDTA 27.5 mM, polybrene 1 mg ml⁻¹) was added (0.25 mM final). The reaction was stopped 2 min later by the addition of a 50% aqueous acetic acid solution and the A₄₀₅ was read on a spectrophotometer. The activity per mg of the compounds was determined by comparison with a calibrated standard.

Affinity for human antithrombin. Affinity of the oligosaccharides for human antithrombin was determined by fluorescence as described²⁶ using a Perkin Elmer LS-50 type spectrofluorometer at a λ excitation of 280 nm and a λ emission of 338 nm at 37 °C, under continuous stirring. Oligosaccharides were added into 2 ml of Tris-HCl buffer (0.01 mM, pH 7.0) containing 0.15 M NaCl and 5–60 nM antithrombin. The ratio and the concentrations of antithrombin–oligosaccharide complex were calculated and dissociation constants (K_d) were determined from Scatchard plots using RS/1 software.

Platelet activation in the presence of plasma from HIT patients. Increasing concentrations of the various compounds were incubated with washed human platelets (3 × 10⁵ platelets per µl) previously incubated with ¹⁴C-serotonin (0.05 µCi ml⁻¹) in the presence of plasma from HIT patients. After 60 min of incubation, platelet activation was measured by ¹⁴C-serotonin release as described²⁷.

Venous and arterial thrombosis in the rat. To determine their venous antithrombotic activity, the compounds were administered intravenously (i.v.) to rats 5 min before venous thrombosis induced in the vena cava by a combination of stasis and tissue factor (1 ng kg⁻¹, i.v.) as described previously¹⁸. Their arterial antithrombotic activity was determined in an arterio-venous shunt model as described^{18,19}. Grouped data were analysed for significance by comparison with the vehicle-treated group using the Mann–Whitney test with Holm–Bonferroni adjustment taking *P* < 0.05 to indicate a significant difference.

Bleeding time in the rat. Increasing doses of the various compounds were administered i.v. to rats 15 s before the transection of the tip of the tail of pentobarbital-anaesthetized rats. Bleeding time was then measured as described¹⁸.

Received 21 December 1998; accepted 5 February 1999.

1. Lane, D. A. & Lindahl, U. (eds) *Heparin* (Arnold, London, 1989).
2. Verstraete, M. Pharmacotherapeutic aspects of unfractionated and low molecular weight heparins. *Drugs* **40**, 498–530 (1990).
3. Warkentin, T. E., Chong, B. H. & Greinacher, A. Heparin-induced thrombocytopenia: Towards consensus. *Thromb. Haemost.* **79**, 1–7 (1998).
4. Thomas, D. P. Does low molecular weight heparin cause less bleeding? *Thromb. Haemost.* **78**, 1422–1425 (1997).
5. Paulsen, H. Advances in selective chemical syntheses of complex oligosaccharides. *Angew. Chem. Int. Ed. Engl.* **21**, 155–173 (1982).
6. Choay, J. et al. Structure–activity relationship in heparin: a synthetic pentasaccharide with high affinity for antithrombin III and eliciting high anti-factor Xa Activity. *Biochem. Biophys. Res. Commun.* **116**, 492 (1983).
7. van Boeckel, C. A. A. & Petitou, M. The unique antithrombin binding domain of heparin: a lead to new synthetic antithrombotics. *Angew. Chem. Int. Ed. Engl.* **32**, 1671–1690 (1993).
8. Casu, B. Structure and biological activity of heparin. *Adv. Carbohydr. Chem. Biochem.* **43**, 51–134 (1985).
9. Olson, S. T. & Björk, I. Regulation of thrombin activity by antithrombin and heparin. *Semin. Thromb. Haemost.* **20**, 373–409 (1994).
10. Grootenhuys, P. D. J., Westerduin, P., Meuleman, D., Petitou, M. & van Boeckel, C. A. A. Rational

- design of synthetic heparin analogues with tailor-made coagulation factor inhibitory activity. *Nature Struct. Biol.* **2**, 736–739 (1995).
11. Laurent, T. C., Tengblad, A., Thunberg, L., Höök, M. & Lindahl, U. The molecular-weight-dependence of the anti-coagulant activity of heparin. *Biochem. J.* **175**, 691–701 (1978).
 12. Oosta, G. M., Gardner, W. T., Beeler, D. L. & Rosenberg, R. D. Multiple functional domains of the heparin molecule. *Proc. Natl Acad. Sci. USA* **78**, 829–833 (1981).
 13. Lane, D. A., Denton, J., Flynn, A. M., Thunberg, L. & Lindahl, U. Anticoagulant activities of heparin oligosaccharides and their neutralization by platelet factor 4. *Biochem. J.* **218**, 725–732 (1984).
 14. Danielsson, A., Raub, E., Lindahl, U. & Björk, I. Role of ternary complexes, in which heparin binds both antithrombin and proteinase, in the acceleration of the reactions between antithrombin and thrombin or factor Xa. *J. Biol. Chem.* **261**, 15467–15473 (1986).
 15. Basten, J. et al. Biologically active heparin-like fragments with a "non-glycosamino" glycan structure. Part 3: O-alkylated-O-sulphated pentasaccharides. *Bioorg. Med. Chem. Lett.* **2**, 905–910 (1992).
 16. Petitou, M. et al. Synthesis and pharmacological properties of a close analogue of an antithrombotic pentasaccharide (SR 90107A/ORG 31540). *J. Med. Chem.* **40**, 1600–1607 (1997).
 17. Petitou, M. et al. First synthetic carboxyhydrates with the full anticoagulant properties of heparin. *Angew. Chem. Int. Ed. Engl.* **37**, 3009–3014 (1998).
 18. Herbert, J.-M. et al. Biochemical and pharmacological properties of SANORG 32701. Comparison with the "synthetic pentasaccharide" (SR 90107A/ORG 31540) and standard heparin. *Circ. Res.* **76**, 590–600 (1996).
 19. Umetsu, T. & Sanai, K. Effect of KC-6141, an anti-aggregating compound, on experimental thrombosis in rats. *Thromb. Haemost.* **39**, 74–83 (1978).
 20. Jin, J. et al. The anticoagulant activation of antithrombin by heparin. *Proc. Natl Acad. Sci. USA* **94**, 14683–14688 (1997).
 21. Westerduin, P. et al. Feasible synthesis and biological properties of six "non-glycosamino" glycan analogues of the antithrombin III binding heparin pentasaccharide. *Bioorg. Med. Chem.* **2**, 1267–1280 (1994).
 22. Stubbs, M. T. & Bode, W. The clot thickens: clues provided by thrombin structure. *Trends Biochem. Sci.* **20**, 23–28 (1995).
 23. Sache, E. et al. Studies on a highly active anticoagulant fraction of high molecular weight isolated from porcine sodium heparin. *Thromb. Res.* **25**, 443–458 (1982).
 24. Olson, S. T., Halvorson, H. R. & Björk, I. Quantitative characterization of the thrombin-heparin interaction. Discrimination between specific and non-specific binding models. *J. Biol. Chem.* **266**, 6342–6352 (1991).
 25. Maccarana, M. & Lindahl, U. Mode of interaction between platelet factor 4 and heparin. *Glycobiology* **3**, 271–277 (1993).
 26. Atha, D. H., Lormeau, J. C., Petitou, M., Rosenberg, R. D. & Choay, J. Contribution of 3-O- and 6-O-sulfated glucosamine residues in the heparin-induced conformational change in antithrombin III. *Biochemistry* **26**, 6454–6461 (1987).
 27. Sheridan, D., Carter, C. & Kelton, J. G. A diagnostic test for heparin-induced thrombocytopenia. *Blood* **67**, 27–30 (1986).

Acknowledgements. This work is part of a collaborative project between N. V. Organon and Sanofi Recherche on antithrombotic oligosaccharides.

Correspondence and requests for materials should be addressed to J.-M.H. (e-mail: jean-marc.herbert@sanofi.com).

β-Catenin regulates expression of cyclin D1 in colon carcinoma cells

Osamu Tetsu & Frank McCormick

University of California, San Francisco, School of Medicine, Cancer Research Institute, San Francisco, California 94143-0128, USA

Mutations in the adenomatous polyposis coli (APC) tumour-suppressor gene occur in most human colon cancers¹. Loss of functional APC protein results in the accumulation of β-catenin². Mutant forms of β-catenin have been discovered in colon cancers that retain wild-type APC genes^{3,4}, and also in melanomas⁵, medulloblastomas⁶, prostate cancer⁷ and gastric⁸ and hepatocellular^{9,10} carcinomas. The accumulation of β-catenin activates genes that are responsive to transcription factors of the TCF/LEF family, with which β-catenin interacts^{11–15}. Here we show that β-catenin activates transcription from the *cyclin D1* promoter, and that sequences within the promoter that are related to consensus TCF/LEF-binding sites are necessary for activation. The oncoprotein p21^{ras} further activates transcription of the *cyclin D1* gene, through sites within the promoter that bind the transcriptional regulators Ets or CREB. Cells expressing mutant β-catenin produce high levels of cyclin D1 messenger RNA and protein constitutively. Furthermore, expression of a dominant-negative form of TCF in colon-cancer cells strongly inhibits expression of cyclin D1 without affecting expression of cyclin D2, cyclin E, or cyclin-dependent kinases 2, 4 or 6. This dominant-

negative TCF causes cells to arrest in the G1 phase of the cell cycle; this phenotype can be rescued by expression of cyclin D1 under the cytomegalovirus promoter. Abnormal levels of β-catenin may therefore contribute to neoplastic transformation by causing accumulation of cyclin D1.

Genes that might be targets for transcriptional activation by the β-catenin/TCF complex have been identified in various organisms. These include the mouse E-cadherin gene¹⁶, the connexin 43 gene¹⁷, *Drosophila ultrabithorax*¹⁸, *Xenopus siamois*¹⁹ and *nodal-related 3* (ref. 20). More recently, the *c-myc* gene has been implicated as a target of the APC pathway in colon cancer²¹. *c-myc* expression is inhibited by expression of APC, and the *c-myc* promoter contains elements that bind TCF *in vitro*. Also, many colon carcinomas show high levels of *c-myc* expression. However, it is not clear whether β-catenin activates *c-myc* directly, nor to what extent *c-myc* expression contributes to malignant growth. To identify candidate target genes of β-catenin-mediated transcription, we searched the GenBank database for human genes involved in controlling cell growth whose promoter region contains the core TCF/LEF-binding site 5'-A/T A/T CAAAG-3' (ref. 14). Through a series of computer searches, we identified four candidate target genes, *cyclin D1*, *cyclin A*, *cdc2* and *cdc25C*. Of these, *cyclin D1* is of particular interest because cyclin D1 is over-expressed in many colon carcinomas, although the gene is rarely amplified²² and inhibition of *cyclin D1* expression causes growth arrest in colon carcinoma cell lines²³. The core TCF/LEF consensus site does not occur in promoter regions of other cyclins, including *cyclin D2* and *cyclin D3*, *cyclin E*, *cyclin G*, nor in other cyclin-dependent kinases whose promoter regions have been sequenced, such as *cdk2*, *cdk4*, *cdk5*, *cdk6* and *cdk7*.

To determine whether β-catenin regulates the transcription of these genes, we transfected luciferase reporter genes into HeLa cells and measured their response to β-catenin. Two forms of β-catenin expression vector were tested. One encodes wild-type β-catenin, the other encodes a mutant form that lacks amino acids 29–48 and

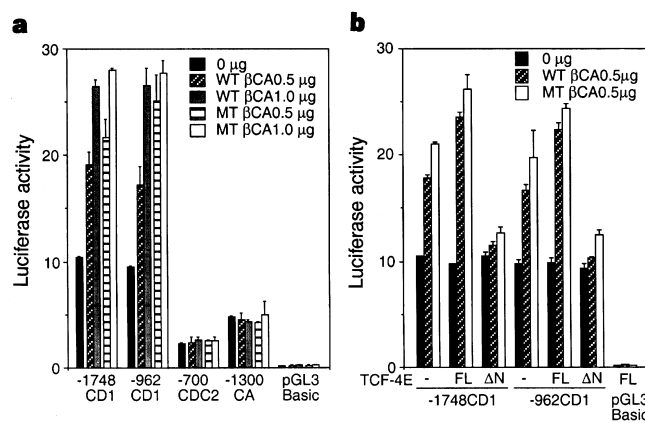


Figure 1 β-Catenin activates the *cyclin D1* promoter. **a**, HeLa cells were transfected with promoters from *cyclin D1*, *cdc2* or *cyclin A* driving luciferase transcription, and 0.5 or 1 μg of wild-type (WT) or mutant-type (MT) β-catenin (βCA) expression plasmids. Two forms of *cyclin D1* promoter were used, one consisting of 1,748 bp from the transcriptional start (–1748CD1), the other of 962 bp (–962CD1). Empty pcDNA3 (Invitrogen) was added to 2 μg of plasmid DNA. 50 ng of pRL-TK *renilla* luciferase reporter construct (Promega) was co-transfected to each sample to normalize transfection efficiency. The ratio of reporter luciferase activity to control *renilla* luciferase activity is indicated. The assay was performed 24 h after transfection. All experiments are expressed as mean ± s.d. of triplicate cultures. **b**, Effects of dominant-negative TCF-4E on β-catenin-dependent *cyclin D1* promoter activity. Full-length (FL) TCF-4E complementary DNA was subcloned into Myc-tagged pcDNA3 from TCF-4E pcDNA1. N-terminal-deleted (ΔN) Myc-tagged TCF-4E pcDNA3 was generated by PCR-based site-directed deletion mutagenesis.

therefore does not have the phosphorylation sites necessary for APC-mediated degradation⁵. Two *cyclin D1* reporter plasmids were tested. One of these (-1748CD1) corresponds to the original fragment of *cyclin D1* 5' sequence cloned from the PRAD1 breakpoint²⁴, and the other, -962CD1, is the minimum 5' sequence that retains responsiveness to p21^{ras} (ref. 25). Figure 1a shows that both -1748CD1 and -962CD1 reporters were strongly transcribed in response to wild-type and mutant β -catenin, in a dose-dependent manner. We obtained similar results using NIH3T3 cells or human embryonic kidney 293 cells (data not shown). In contrast, we observed no enhanced transcription in cells transfected with either -1,300 *cyclin A* or -700 *cdc2* reporter constructs. Similarly, no activation was seen with -7.5-kilobase (kb) *cyclin A* or -6.7-kb *cdc2* reporter genes (data not shown). To determine whether activation of the *cyclin D1* promoter by β -catenin depends on TCF, we co-transfected a dominant-negative TCF-4E (Fig. 1b). We found that dominant-negative TCF-4E strongly inhibited activation by β -catenin, but did not affect the basal level of transcription. Full-length TCF-4E increased activation slightly; we presume this effect is modest because HeLa cells already express sufficient levels of TCF to support β -catenin-dependent activation. Indeed,

we have used the polymerase chain reaction with reverse transcription (RT-PCR) to show that HeLa cells do indeed express TCF-4E mRNA at levels comparable to SW480 cells, in which this factor was first detected (data not shown).

The consensus TCF-binding site mapped at 25 base pairs (bp) 5' of the CREB-binding site within the *cyclin D1* promoter. The -962CD1 promoter fragment also has three potential Ets-binding sites and one AP-1 site (Fig. 2a). These factors have been shown previously to collaborate with LEF-1 to regulate the transcription of TCR- α (ref. 15). To determine whether these elements are necessary for the β -catenin responsiveness of the *cyclin D1* promoter, mutations were introduced into each region. Elimination of the Ets-B or CREB sites reduced overall transcriptional activity of the -962CD1 reporter gene, but did not affect its response to β -catenin. To our surprise, elimination of the consensus TCF (1) site did not eliminate β -catenin responsiveness completely, although the effect was significantly diminished (Fig. 2b). We therefore searched the *cyclin D1* promoter for other potential TCF-binding sites that differed slightly from the core consensus. Three additional sequences, TCF (2), TCF (3) and TCF (4), were identified close to the perfect TCF consensus site TCF (1), and another sequence, TCF (0), was identified far from

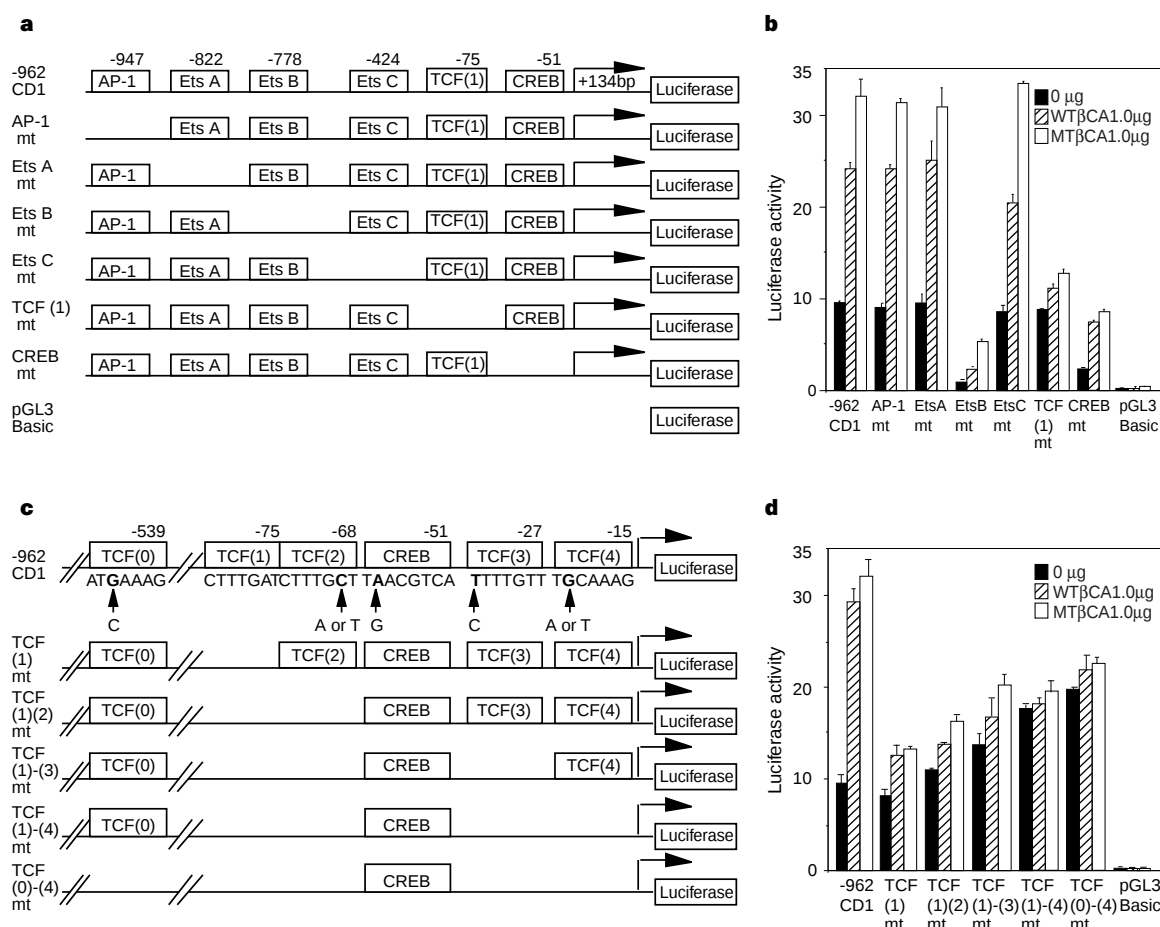


Figure 2 β -catenin-responsive elements within the *cyclin D1* promoter. **a**, Schematic representation of -962 *cyclin D1* (CD1) pGL3 basic luciferase reporter constructs and deletion mutants. Possible transcription-factor-binding sites are indicated. The nucleotide sequence of this region is also available from the NCBI database (accession no. L09054). AP-1, Ets (A, B, C), TCF (1) and CREB sites were eliminated as shown. All mutations were checked by sequencing. **b**, Response of mutant promoters to β -catenin. Reporter-gene assay with -962 *cyclin D1* and its derivative in HeLa cells. Cells were transfected with 1 μ g of -962CD1 AP-1 mutant, Ets-A mutant, Ets-B mutant, Ets-C mutant, TCF (1) mutant, CREB mutant or empty pGL3 basic luciferase plasmids, and 1 μ g WT or MT β -catenin (β CA), pcDNA3. The ratio of reporter luciferase activity to control *renilla* luciferase activity is

indicated, as in Fig. 1. **c**, Schematic representation of -962CD1 luciferase reporter constructs and mutant reporter constructs, indicating potential TCF-binding sites, with their sequences. TCF(0), TCF(2), TCF(3) and TCF(4) diverge by one nucleotide from the perfect TCF consensus, as indicated with arrows. Site-directed deletions for TCF(1)(2), (1)-(3), (1)-(4), and (0)-(4) were checked by sequencing. **d**, Reporter-gene assay with -962CD1 and its potential TCF-binding sites mutants in HeLa cells. Cells were transfected with 1 μ g of either -962CD1, TCF(1) mutant, TCF(1)(2) mutant, TCF(1)-(3) mutant, TCF(1)-(4) mutant, TCF(0)-(4) or control pGL3 basic plasmids, and 1 μ g of either WT or MT β -catenin (β CA) pcDNA3 expression vector. 50 ng pRL-TK reporter construct was co-transfected to normalize for transfection efficiency, as in **b**.

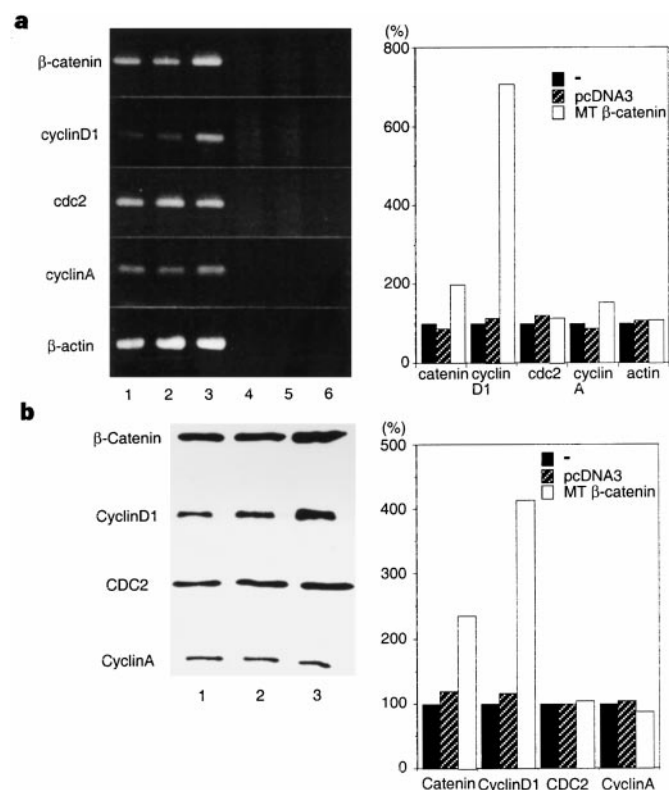


Figure 3 Expression of β -catenin, cyclinD1, cdc2 and cyclinA in HeLa cells expressing mutant β -catenin. **a**, Left, Quantitative RT-PCR analysis of non-transfected HeLa cells (lanes 1, 4), empty pcDNA3-transfected HeLa cells (lanes 2, 5) and MT β -catenin-transfected stable HeLa cells (lanes 3, 6). Lanes 1–3 are with reverse transcriptase; lanes 4–6 are without reverse transcriptase. 50 ng of total RNA-derived cDNA was used for PCR reactions. Right, Quantitation of data is shown on the left. Expression levels relative to control HeLa cells are shown. **b**, Left, western blot analysis of non-transfected HeLa cells (lane 1), empty pcDNA3-transfected stable HeLa cells (lane 2) and MT β -catenin-transfected stable HeLa cells (lane 3), 10 μ g of protein was loaded. Right, Quantified representation of western blot data. Expression levels relative to control HeLa cells are shown.

these sites at –539 (Fig. 2c). Figure 2d shows that deletion of increasing numbers of putative TCF-binding sites resulted in an increase in the basal level of reporter activity. This is consistent with previous analysis of the *siamois* promoter of *Xenopus*¹⁸ and the *ultrabithorax* midgut enhancer of *Drosophila*¹⁹, and confirms that TCF acts as a negative regulator of transcription in the absence of β -catenin. In contrast, deletion of increasing numbers of TCF sites results in a gradual reduction in responsiveness to β -catenin. We conclude from these analyses that the responsiveness of the *cyclin D1* promoter to β -catenin is dependent on TCF sites, but is independent of a specific context of other transcription factors.

To investigate the effect of β -catenin on endogenous *cyclin D1* expression, we generated stable HeLa cell cultures expressing mutant β -catenin. HeLa cells were chosen because they express TCF-4E, and because they lack a functional checkpoint for the retinoblastoma (Rb) tumour suppressor. Their growth should therefore be unaffected by high levels of cyclinD1 protein. Figure 3 shows that HeLa cells expressing mutant β -catenin do indeed express high levels of cyclin D1 mRNA constitutively (about seven times higher than control cells) and produce about four times more cyclin D1 protein. In contrast, *cdc2* expression was approximately the same in transfected cells as in controls. Cyclin A mRNA was upregulated by about 50%, but the level of cyclin A protein was unchanged (Fig. 3b).

In the progression of colon cancer from normal epithelium to full malignancy, Ras mutations are frequently involved, and occur later than mutations in APC and, presumably, β -catenin¹. Ras has been shown previously to activate *cyclin D1* expression²⁵. We therefore tested the possibility that Ras and β -catenin act together to increase *cyclin D1* transcription. We co-transfected HeLa cells with the combination of β -catenin and valine 12 p21^{ras}, with both the –1,748 and –962 *cyclin D1* reporter constructs. Valine 12 p21^{ras} increased the response of both promoters to β -catenin, and had a similar effect on its own (Fig. 4a). Figure 4b shows that deletion of the Ets-B- or CREB-binding site strongly inhibited Ras-mediated transcriptional enhancement of the *cyclin D1* promoter. Deletion of TCF sites, which prevented β -catenin-mediated activation, did not affect Ras-mediated enhancement, even though the basal level of activity was higher than in the –962CD1 reporter (Fig. 4c). We conclude from these data that Ras activation of the *cyclin D1*

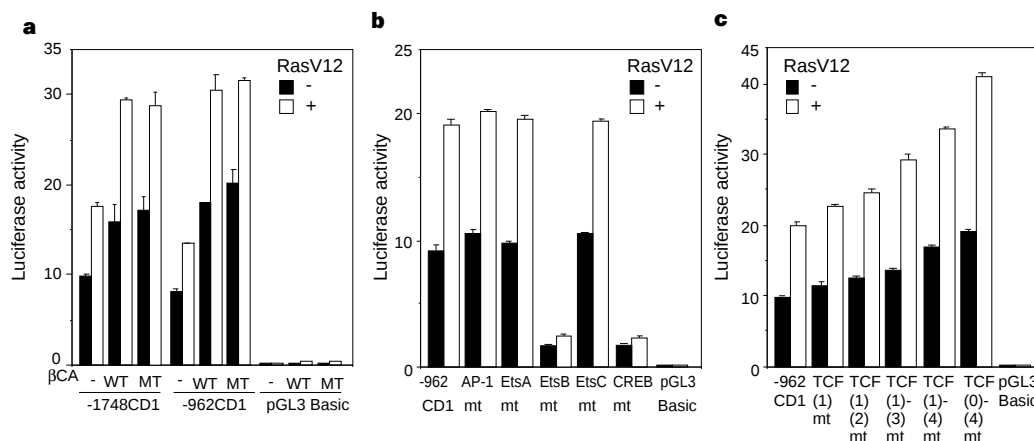


Figure 4 Effects of valine 12 Ras on transcription from the *cyclin D1* promoter. **a**, Reporter-gene assays with –1748CD1 and –962CD1 pGL3 Basic in HeLa cells. Cells were transfected with 1 μ g of –1748CD1 or –962CD1 or control pGL3 basic plasmids co-transfected in the combination of 0.5 μ g wild-type (WT) or mutant-type (MT) β -catenin (β CA), and 0.5 μ g valine 12 p21^{ras} mutant pcDNA3 (Ras V12) expression vectors. 50 ng pRL-TK reporter construct was co-transfected to normalize for transfection efficiency. **b**, Reporter-gene assays with –962CD1 and its derivative mutants, AP-1, Ets (A, B, C) and CREB, in HeLa cells. Cells were

transfected with 1 μ g of –962CD1 or AP-1 or Ets (A, B, C) or CREB or control empty pGL3 Basic plasmids co-transfected valine 12 p21^{ras} pcDNA3 (Ras V12) expression vectors. **c**, Reporter-gene assay with –962CD1 and its potential TCF-binding site mutants indicated in Fig. 2c, in HeLa cells. Cells were transfected with 1 μ g of either –962CD1, TCF(1) mutant, TCF(1)(2) mutant, TCF(1)–(3) mutant, TCF(1)–(4) mutant, TCF(0)–(4) or control pGL3 Basic plasmids, and 1 μ g valine 12 p21^{ras} pcDNA3 (Ras V12) expression vector.

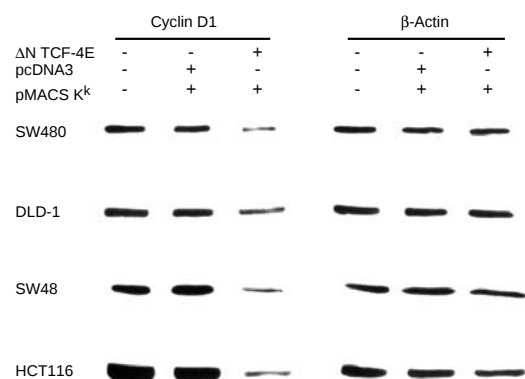


Figure 5 Repression of cyclin D1 expression by a dominant-negative TCF-4E. Cycling colon cancer cells were transfected with either 7.5 μg pcDNA3 or ΔN TCF-4E pcDNA3 expression vector with 2.5 μg pMACS K^k (Miltenti Biotech). 48 h after transfection, cells expressing the truncated H-2K^k were collected by beads coated with an antibody to mouse H-2K^k according to its instruction manual. Total cellular protein was isolated immediately and subjected to immunoblotting with cyclin D1 monoclonal antibody. Subsequently, the identical membranes were blotted with β-actin to indicate that equal amounts of protein were loaded in each lane.

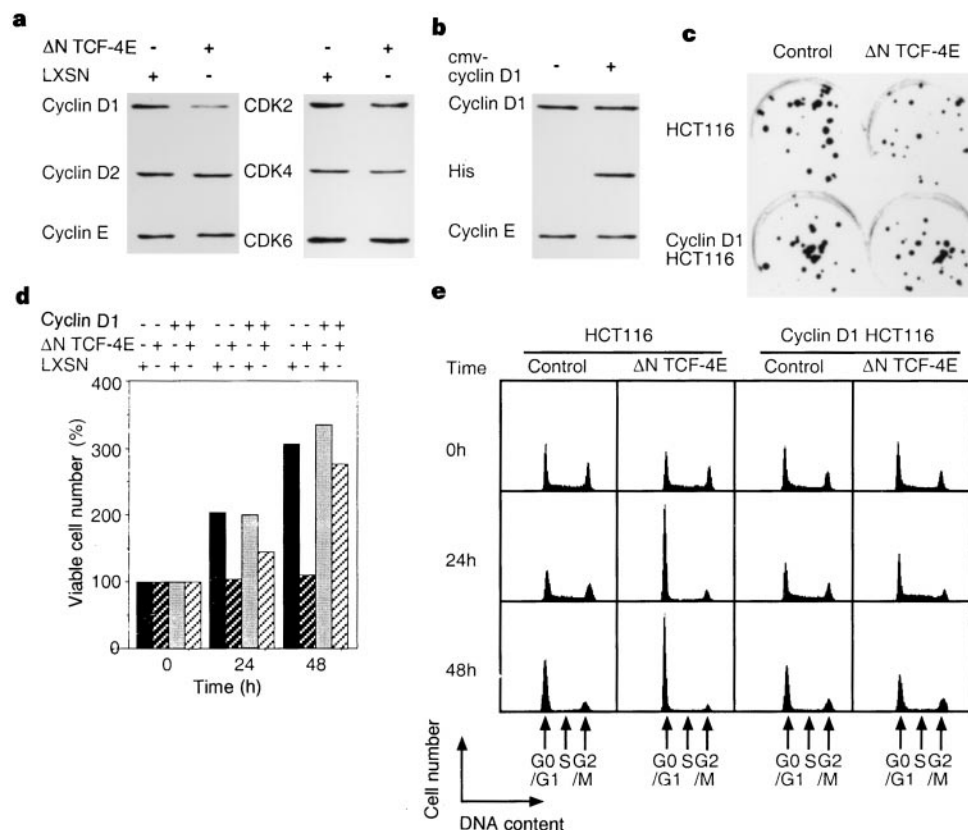


Figure 6 G1 growth arrest in HCT116 colon cancer cells by dominant negative TCF-4E, and rescue by ectopic expression of cyclin D1. Cell-cycle analysis was performed after ΔN TCF-4E retrovirus infection of HCT116 and of HCT116 cells transfected with His-tagged *cyclin D1* cDNA (Invitrogen) expressed from the cytomegalovirus promoter (Cyclin D1 HCT116). Exponentially growing cells were infected with LXSN (Clontech), empty vector retrovirus (control) or retrovirus expressing ΔN TCF-4E. **a**, Western blot analysis of cells infected with dominant-negative TCF-4E. Cell extracts were prepared 24 h after infection and the blot was first probed with a monoclonal antibody against cyclin D1. The blot was then probed with antibodies against cyclins D2, D3 and E and cdk2 and cdk4 and cdk6. Cyclin D3 was not detectable in these extracts. **b**, Western blot of HCT116 cells expressing *cyclin D1* from the cytomegalovirus (cmv) promoter. Extracts were

promoter in HeLa cells requires both the Ets-B- and CREB-binding sites and, as discussed above, is independent of the TCF sites.

Cyclin D1 protein is highly expressed in patients with adenomatous polyps, primary colorectal adenocarcinoma²² and familial adenomatous polyposis, and in mice bearing intestinal adenomas from multiple intestinal neoplasia²⁶. To investigate the possibility that β-catenin contributes to elevated *cyclin D1* expression, we determined whether endogenous *cyclin D1* expression in colon cancers was inhibited by transient expression of dominant-negative TCF-4E. We tested four colon cancer cell lines. Two of these, SW 480 and DLD-1, express mutant APC but wild-type β-catenin; the other two cell lines, SW 48 and HCT 116, contain wild-type APC but mutant β-catenin. Figure 5 shows that, in each cell line, dominant-negative TCF-4E caused a significant reduction in the levels of cyclin D1 protein. We then tested whether reduced cyclin D1 caused by disrupting signalling from β-catenin to TCF would result in growth arrest. This effect was predicted because cyclin D1 has been shown previously to be rate limiting for progression through the G1 phase of the cell cycle²⁷. In addition, we tested whether HCT116 cells expressing *cyclin D1* from the constitutive cytomegalovirus promoter would be resistant to growth-inhibitory effects caused by blocking signalling from β-catenin to TCF. HCT116 cells were infected with a retrovirus expressing dominant-negative TCF-4E. These cells express an

prepared from exponentially growing HCT116 cells or cells expressing *cyclin D1* from the cytomegalovirus promoter (cmv cyclic D1), and blotted with antibodies specific for cyclin D1, His-tag (Invitrogen) or cyclin E. **c**, Colony formation assay on cells infected with dominant-negative TCF-4E and selected with G418. 50 single cells from HCT116 and Cyclin D1-HCT116 were infected with LXSN or ΔN TCF-4E LXSN, respectively. These vectors express the neomycin gene. 14 days after infection and growth in G418, cells were stained with 0.5% crystal violet containing 20% ethanol. **d**, Viable cell number after infection. Cells were collected at the indicated times (0, 24, 48 h) after infection, stained with trypan blue and counted with a haemocytometer. **e**, Flow cytometric cell-cycle analysis. Cells were stained with propidium iodide to show cell-cycle distribution by DNA content. 10,000 cells were tested. G0/G1, S and G2/M phases of the cell cycle are indicated by arrows.

intact Rb protein, and so their growth should be sensitive to levels of *cyclin D1* expression. Figure 6 shows that levels of cyclin D1 are decreased in infected cells, as expected from the results described in Fig. 5, and that levels of cyclin D2, cyclin E, cdk2, cdk4 and cdk6 are unaffected (Fig. 6a). Cyclin D3 was not detected in these cells. Figure 6b shows levels of cyclin D1 in HCT116 cells expressing His-tagged cyclin D1 from the cytomegalovirus promoter: total levels of cyclin D1 are not affected significantly by ectopic expression from this promoter.

Expression of dominant-negative TCF-4E in HCT116 cells caused severe growth arrest, as measured in a colony-forming assay, or by counting viable cells. In the colony-forming assay, cells expressing the dominant-negative TCF virus were selected in G418 and, after 2 weeks, the colonies of cells were stained with crystal violet. The number of colonies formed in the presence of dominant-negative TCF-4E was similar to controls, but the colonies were significantly smaller (Fig. 6c). This suggests an effect on cell division without an effect on plating efficiency or viability. This was confirmed by counting viable cells after infection: no change in the number of viable cells was measured over 48 h; in this time, control cells had more than doubled (Fig. 6d). Cells arrested by expression of dominant-negative TCF-4E accumulated in the G1 phase of the cell cycle (Fig. 6e). No sub-G1 population was detected, indicating no detectable apoptosis. Ectopic expression of cyclin D1 rescued cells from this G1 block, resulting in an increase in colony size and viable cell number (Fig. 6c, d) and restoration of a normal cell cycle (Fig. 6e). These results indicate that reduced levels of cyclin D1 were indeed responsible for growth inhibition. Growth of HCT116 cells is therefore strongly dependent on β -catenin/TCF-mediated expression of *cyclin D1*. Furthermore, this effect appears to be direct, as levels of other G1 cyclins and their associated kinases, none of which contains TCF consensus sites in their promoters, were not affected during this time (Fig. 6a). The lack of effect on cyclin D2 was particularly striking since this protein resembles cyclin D1 in many of its biological activities^{27,28}.

We show here that β -catenin activates the transcription of *cyclin D1* through TCF-binding sites within the promoter. Activated Ras can further increase transcription of *cyclin D1* through independent signals that require Ets- and CREB-binding sites within the promoter. Expression of *cyclin D1* is strongly dependent on β -catenin/TCF and has a direct effect on cell proliferation. Increased transcription of *cyclin D1* by Ras and β -catenin is likely to play an important role in the development of colon cancer and other malignancies involving these pathways. □

Methods

Transient reporter assays. Promoter fragments were subcloned into pGL3 Basic luciferase reporter plasmid (Promega). FuGENE 6 (Boehringer) was used for transfections. Reporter gene assays were performed with the Dual-luciferase reporter assay system (Promega). 50 ng pRL-TK (Promega) *renilla* luciferase was co-transfected in each sample as an internal control for transfection efficiency. All experiments were performed in triplicate from independent cell cultures. *cyclin D1*, *cdc2* and *cyclin A* reporter constructs were gifts from R. Pestell, Albert Einstein College of Medicine, New York, T. Born and J. Feramisco, University of California at San Diego, and B. Henglein, Institut Curie, Paris, respectively. TCF-4E pcDNA1 and Myc-tagged pcDNA3 were gifts from H. Clevers, University Hospital, Utrecht, and B. Rubinfeld, Onyx Pharmaceuticals, respectively.

Site-directed deletion mutagenesis. This was performed with the Exsite PCR-based site-directed mutagenesis kit (Stratagene). Details of PCR primers can be obtained from the authors (e-mail: tetsu@cc.ucsf.edu). To mutate specific transcription-factor-binding sites, the following sequences were removed: TCF (1), CTTTGAT; AP-1, TGAGTCA; Ets-A, AGGAA; Ets-B, AGGAA, Ets-C, TTCCT; CREB, TAACGTCA. All mutations were checked by sequencing. Myc-epitope-tagged wild-type β -catenin pcDNA3 expression vector was a gift from P. Polakis, Onyx Pharmaceuticals.

RT-PCR. Total RNA was prepared using Trizol (Gibco). RT-PCR was performed

using a RAP-PCR kit (Stratagene). PCR reactions used AmpliTaq DNA polymerase (Perkin Elmer). Cycling conditions were 5 min at 94 °C, 30 cycles at 94 °C for 1 min, 55 °C for 1 min, 72 °C for 1 min.

Western blot analysis. Protein was prepared with cell lysis buffer consisting of 10 mM Tris-HCl (pH 7.4), 25 mM NaCl and 2 mM DTT for β -catenin to obtain cytosol fractions, or 50 mM Tris-HCl (pH 8.0), 120 mM NaCl, 0.25% NP-40 and 0.1% SDS for cyclin D1, cdc2, cyclin A to obtain whole cell lysates. Western blots were developed by enhanced chemiluminescence (Amersham). Primary monoclonal and secondary conjugated antibodies were against β -catenin and cdc2 (Transduction), cyclins A, D1, D2, D3 and E, cdk2, 4 and 6 (Santa Cruz). Goat anti-mouse IgG peroxidase (Pierce) and donkey anti-rabbit IgG peroxidase (Amersham) were used. Quantitation was by Gel Doc 1000 using its Molecular Analyst software (Bio-Rad).

Cell-cycle analysis. Cells were pelleted and resuspended in 1 ml of 0.1% sodium citrate containing 0.3% NP-40, 0.002 mg ml⁻¹ RNase and 50 mg ml⁻¹ propidium iodide, and incubated for 30 min on ice. The profile of cells in the G0/G1, S and G2/M phases of the cell cycle were analysed by the UCSF Cancer Center Cytometry Core, on a FACSCaliber with Cellquest Software (Becton Dickinson).

Received 30 December 1998; accepted 17 February 1999.

- Kinzler, K. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).
- Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B. & Polakis, P. Regulation of intracellular β -catenin levels by the adenomatous polyposis coli (APC) tumor-suppressor protein. *Proc. Natl Acad. Sci. USA* **92**, 3046–3050 (1995).
- Korinek, V. *et al.* Constitutive transcriptional activation by a β -catenin-Tcf complex in APC-/- colon carcinoma. *Science* **275**, 1784–1787 (1997).
- Morin, P. *et al.* Activation of β -catenin-Tcf signaling in colon cancer by mutations in β -catenin or APC. *Science* **275**, 1787–1790 (1997).
- Rubinfeld, B. *et al.* Stabilization of β -catenin by genetic defects in melanoma cell lines. *Science* **275**, 1790–1792 (1997).
- Zurawel, R., Chiappa, S., Allen, C. & Raffel, C. Sporadic medulloblastomas contain β -catenin mutations. *Cancer Res.* **58**, 896–899 (1998).
- Voeller, H., Truica, C. & Gelmann, E. β -catenin mutations in human prostate cancer. *Cancer Res.* **58**, 2520–2523 (1998).
- Kawashiri, J. *et al.* Loss of E-cadherin-dependent cell-cell adhesion due to a mutation of the β -catenin gene in a human cancer cell line, HSC-39. *Mol. Cell. Biol.* **15**, 1175–1181 (1995).
- Miyoshi, Y. *et al.* Activation of the β -catenin gene in primary hepatocellular carcinomas by somatic alterations involving Exon 3. *Cancer Res.* **58**, 2524–2527 (1998).
- De La Coste, A. *et al.* Somatic mutations of the β -catenin gene are frequent in mouse and human hepatocellular carcinomas. *Proc. Natl Acad. Sci. USA* **95**, 8847–8851 (1998).
- Molenaar, M. *et al.* XTcf-3 transcription factor mediates β -catenin-induced axis formation in *Xenopus* embryos. *Cell* **86**, 391–399 (1996).
- Behrens, J. *et al.* Functional interaction of β -catenin with the transcription factor LEF-1. *Nature* **382**, 638–642 (1996).
- Clevers, H. & Grosschedl, R. Transcriptional control of lymphoid development: lessons from gene targeting. *Immunol. Today* **17**, 336–343 (1996).
- van de Wetering, M. *et al.* Armadillo coactivates transcription driven by the product of the *Drosophila* segment polarity gene *disfused*. *Cell* **88**, 789–799 (1997).
- Geise, K., Kingsley, C., Kirshner, J. & Grosschedl, R. Assembly and function of a TCR- α enhancer complex is dependent on LEF-1-induced DNA bending and multiple protein-protein interactions. *Genes Dev.* **9**, 995–1008 (1995).
- Huber, O. *et al.* Nuclear localization of β -catenin by interaction with transcription factor LEF-1. *Mech. Dev.* **59**, 310 (1996).
- van der Heyden, M. *et al.* Identification of connexin43 as a target for wnt signaling. *J. Cell Sci.* **111**, 1741–1749 (1998).
- Riese, J. *et al.* LEF-1, a nuclear factor coordinating signaling inputs from *wingless* and *decapentaplegic*. *Cell* **88**, 777–787 (1997).
- Brannon, M., Gomperts, M., Sumoy, L., Moon, R. & Kimelman, D. A β -catenin/XTcf-3 complex binds to the *siamois* promoter to regulate dorsal axis specification in *Xenopus*. *Genes Dev.* **11**, 2359–2370 (1997).
- McKendry, R., Hsu, S., Harland, R. & Grosschedl, R. LEF-1/TCF proteins mediate Wnt-inducible transcription from the *Xenopus* Nodal related 3 promoter. *Dev. Biol.* **192**, 420–431 (1997).
- He, T. *et al.* Identification of c-myc as a target of the APC pathway. *Science* **281**, 1509–1512 (1998).
- Arber, N. *et al.* Increased expression of cyclin D1 is an early event in multistage colorectal carcinogenesis. *Gastroenterology* **110**, 669–674 (1996).
- Arber, N. *et al.* Antisense to cyclin D1 inhibits the growth and tumorigenicity of human colon cancer cells. *Cancer Res.* **57**, 1569–1574 (1997).
- Motokura, T. & Arnold, A. PRAD1/cyclin D1 proto-oncogene: genomic organization, 5' DNA sequence, and sequence of a tumor-specific rearrangement breakpoint. *Genes Chromosom. Cancer* **7**, 89–95 (1993).
- Albanese, C. *et al.* Transforming p21ras mutants and c-ets-2 activate the cyclin D1 promoter through distinguishable regions. *J. Biol. Chem.* **270**, 23589–23597 (1995).
- Zhang, T. *et al.* Concurrent overexpression of cyclin D1 and cyclin-dependent kinase 4 (cdk4) in intestinal adenomas from multiple intestinal neoplasia (Min) mice and human familial adenomatous polyposis patients. *Cancer Res.* **57**, 169–175 (1997).
- Quelle, D. *et al.* Overexpression of mouse D-type cyclins accelerates G1 phase in rodent fibroblasts. *Genes Dev.* **7**, 1559–1571 (1993).
- Lahti, J. *et al.* Elimination of cyclin D1 in vertebrate cells leads to an altered cell cycle phenotype, which is rescued by overexpression of murine cyclins D1, D2, or D3 but not by a mutant cyclin D1. *J. Biol. Chem.* **272**, 10859–10869 (1997).

Acknowledgements. We thank J. Xie, G. Yount, P. Sabbatini and P. Rodriguez-Viciana for technical advice, H. Clevers for suggestions, and other members of the McCormick and Stokoe labs for useful discussions. This work was supported in part by the Daiichi Cancer Research Program.

Correspondence and requests for materials should be addressed to F.M. (e-mail: mccormick@cc.ucsf.edu).

Local inhibition and long-range enhancement of Dpp signal transduction by Sog

Hilary L. Ashe & Michael Levine

Department of Molecular and Cellular Biology, Division of Genetics,
401 Barker Hall, University of California, Berkeley, California 94720, USA

Extracellular gradients of signalling molecules can specify different thresholds of gene activity in development. A gradient of Decapentaplegic (Dpp) activity subdivides the dorsal ectoderm of the *Drosophila* embryo into amnioserosa and dorsal epidermis^{1,2}. The proteins Short gastrulation³ (Sog) and Tollid⁴ (Tld) are required to shape this gradient. Sog has been proposed to form an inhibitory complex with either Dpp⁵ or the related ligand Screw^{6,7}, and is subsequently processed by the protease Tld⁵. Paradoxically, Sog appears to be required for amnioserosa formation⁸, which is specified by peak Dpp signalling activity^{1,2}. Here we show that the misexpression of *sog* using the *even-skipped* stripe-2 enhancer⁹ redistributes Dpp signalling in a mutant background in which *dpp* is expressed throughout the embryo. Dpp activity is diminished near the Sog stripe and peak Dpp signalling is detected far from this stripe. However, a tethered form of Sog suppresses local Dpp activity without augmenting Dpp activity at a distance, indicating that diffusion of Sog may be required for enhanced Dpp activity and consequent amnioserosa formation. The long-distance stimulation of Dpp activity by Sog requires Tld, whereas Sog-mediated inhibition of Dpp does not. The heterologous Dpp inhibitor Noggin¹⁰ inhibits Dpp signalling but fails to augment Dpp activity. These results suggest an unusual strategy

for generating a gradient threshold of growth-factor activity, whereby Sog and its protease specify peak Dpp signalling far from a localized source of Sog.

We used the *even-skipped* (*eve*) stripe-2 enhancer and the Flp-FRT system⁹ to misexpress *sog* in the *Drosophila* embryo. We visualized *sog* expression using a digoxigenin-labelled antisense RNA probe¹¹. In addition to the endogenous ventrolateral stripes³ (Fig. 1a), we detected an ectopic segmentation stripe of *sog* in transgenic embryos (Fig. 1b). We used *Race* as a marker gene¹² to detect the presumptive amnioserosa and peak Dpp signalling activity¹³. The ectopic *sog* stripe created a gap in the *Race* expression pattern (Fig. 1d; compare with Fig. 1c), consistent with previous evidence that Sog inhibits Dpp signalling by binding either Dpp⁵ or another signalling molecule of the transforming growth factor- β (TGF- β) family, Screw^{6,7}, which is uniformly expressed throughout early embryos¹⁴.

Sog appears to be required for peak Dpp/Screw activity, as *sog* mutants lack amnioserosa⁸. Several amnioserosa marker genes, including *Kruppel*, *rhomboid* and *hindsight*, exhibit broadened patterns of expression that gradually diminish in older embryos^{3,5}. In contrast, the *Race* pattern was not transiently expanded in *sog* mutants: instead, expression was nearly lost in central regions by the onset of gastrulation (Fig. 1e). *Race* may represent a more definitive marker for the presumptive amnioserosa than the genes used in previous studies^{3,5}. However, the anterior *Race* pattern (Fig. 1e) did not coincide with the amnioserosa, but was associated with head structures such as the optic lobe. The anterior pattern was lost in *dpp* mutants¹² but retained in *sog*⁻ embryos (Fig. 1e), indicating that there may be higher levels of Dpp and/or Screw activity in the presumptive head.

The stripe-2-*sog* transgene created a gap in the *Race* pattern, but did not enhance expression above that directed by the endogenous ventrolateral *sog* stripes (Fig. 1d). We introduced the stripe-2-*sog* transgene into *sog*⁻ mutant embryos to determine whether it could

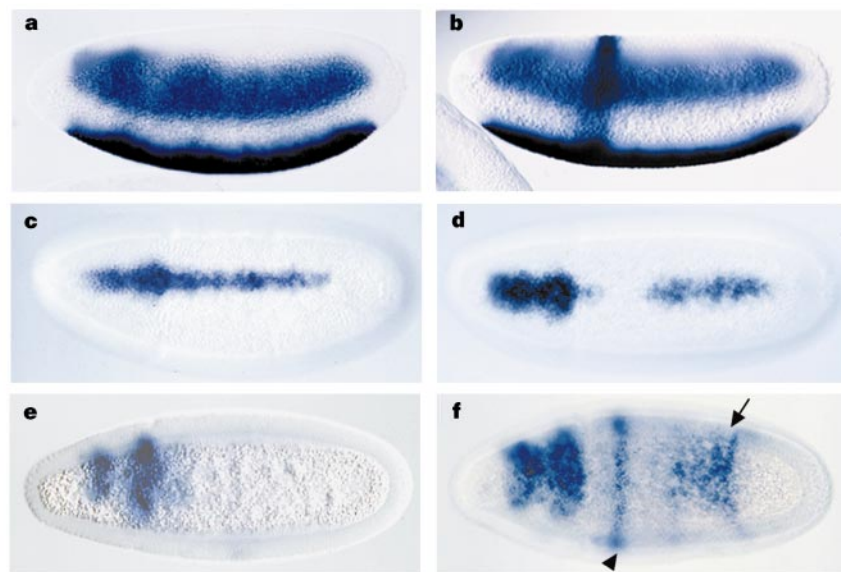


Figure 1 Expression of the stripe-2-*sog* transgene in wild-type and *sog*⁻ embryos. Embryos are orientated with anterior to the left. They were hybridized with digoxigenin-labelled *sog* and/or *Race* antisense RNA probes and histochemically stained with anti-digoxigenin antibodies. **a**, Lateral view of wild-type, late-nuclear-cleavage, cycle-14 embryo stained with a *sog* hybridization probe. *sog* transcripts are detected in two ventrolateral stripes within the presumptive neurogenic ectoderm. **b**, As for **a**, except that the embryo contains the stripe-2-*sog* transgene, which directs an ectopic stripe of *sog* expression. **c**, Dorsal view of a wild-type gastrulating embryo stained with a *Race* hybridization probe. Staining is detected dorsal regions that will form the amnioserosa and head structures. **d**, As for **c**, except that the embryo contains a copy of the stripe-2-*sog*

transgene. There is a gap in the *Race* pattern near the ectopic *sog* stripe. A similar gap was obtained with a modified stripe-2-*sog* transgene that contains the CD2 transmembrane protein (data not shown, and Fig. 4). **e**, Dorsal view of a *sog*⁻ embryo stained with the *Race* hybridization probe. Staining is restricted to anterior regions that will form derivatives of head. There is a loss of staining in central regions that form the amnioserosa. Note that the anterior pattern is expanded into dorsolateral regions as compared with wild-type (**c**). **f**, Dorsal view of a *sog*⁻ embryo that contains the stripe-2-*sog* transgene. The embryo was stained with a mixture of *Race* and *sog* RNA probes. There is a gap in the *Race* pattern centred around the *sog* stripe (arrowhead). *Race* expression is upregulated in posterior regions far from the stripe (arrow).

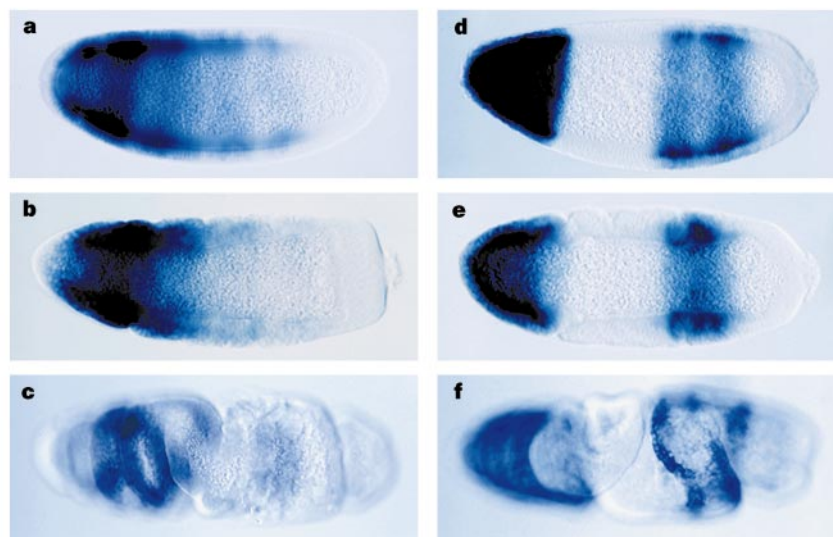


Figure 2 The stripe-2-*sog* transgene induces *Race* expression in *gd*⁻ mutants. Anterior is to the left. Embryos were collected from *gd*⁻ homozygous females and stained with a *Race* hybridization probe. **a**, Cellularized embryo. *Race* transcripts are distributed in a broad anteroposterior gradient, with peak levels in anterior regions. **b**, Gastrulating embryo. *Race* transcripts are gradually lost from posterior regions. **c**, Advanced-stage, elongated embryo. *Race* transcripts are restricted to anterior regions. Weak staining also appears in the anterior and posterior midgut. The mutant embryo has very few amnioserosa cells. **d**, As for **a**, except that the embryo contains the stripe-2-*sog* transgene. There is a loss of *Race* staining in a broad domain on both sides of the *sog* stripe. There is enhanced expression far

from the stripe in both anterior and posterior regions. **e**, As for **b**, except that the embryo contains the stripe-2-*sog* transgene. There is a sharp domain of *Race* repression followed by augmented staining in both anterior and posterior regions. This is particularly evident for the posterior pattern. There is virtually no *Race* expression in this region in *gd*⁻ mutants lacking the *sog* transgene (**b**). **f**, As for **c**, except that the embryo contains the stripe-2-*sog* transgene. The *Race* pattern remains off in central regions, but is sustained in posterior regions. The latter staining pattern coincides with amnioserosa tissue, which is largely absent in embryos lacking the stripe-2-*sog* transgene (**c**).

enhance *Race* expression (Fig. 1f). The transgene provided the only source of Sog (arrowhead), and *Race* expression was now detected in posterior regions far from the stripe (Fig. 1f, arrow); this staining was not observed in mutants lacking the transgene (Fig. 1e). We further investigated the ability of Sog both to inhibit and augment

Dpp/Screw signalling by studying mutants lacking the Dorsal protein nuclear gradient.

The *gastrulation defective* (*gd*) gene encodes a serine protease required for the activation of the Toll signalling pathway¹⁵, and mutant embryos derived from females homozygous for a null

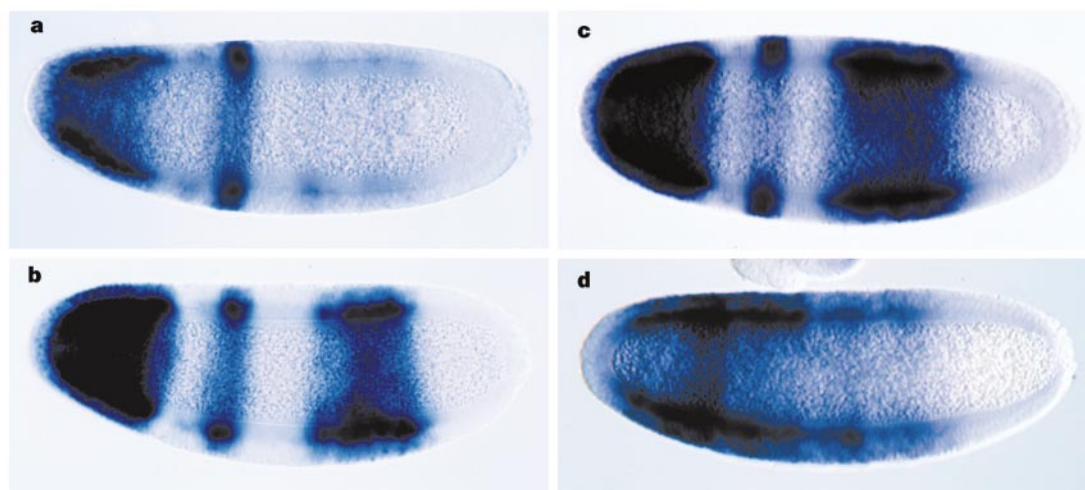


Figure 3 Altering the dose of *dpp* changes the Dpp/Screw signalling pattern. The *gd*⁻ embryos used carried either 1 (**a**), 2 (**b**), or 3 (**c**, **d**) copies of the *dpp*⁺ gene. They were hybridized either with a mixture of *sog* and *Race* RNA probes (**a**–**c**) or with the *Race* probe alone (**d**). **a**, Cellularized *gd*⁻ embryo carrying one copy of *dpp* (*dpp*^{Hin37}/+). *Race* staining is restricted to anterior regions, indicating that the stripe-2-*sog* transgene inhibits Dpp/Screw signalling. However, there is little enhancement in the posterior *Race* pattern (compare with **b**). **b**, Cellularized *gd*⁻ embryo carrying two normal copies of *dpp*. As shown in Fig. 2, there is a gap in the *Race* pattern centred around the *sog* stripe. Enhanced *Race* staining is detected in both anterior and posterior regions. Note that the *sog* stripe is asymmetrically

positioned between the two domains of *Race* expression. This might reflect increased activities of Dpp and/or Screw in anterior regions, thereby requiring higher levels of Sog for inhibition. **c**, Cellularized *gd*⁻ embryo carrying three copies of *dpp*. The zone of *Race* inhibition is narrower than that observed in embryos carrying two copies of *dpp*. Moreover, the *sog* stripe is now positioned symmetrically between the two *Race* domains (compare with **b**), indicating that higher levels of Sog may be required to inhibit the increased levels of Dpp. **d**, As for **c**, except that the embryo lacks the stripe-2-*sog* transgene. *Race* staining is generally increased in these embryos as compared with those containing two copies of *dpp* (Fig. 2a).

mutation in *gd* lack a Dorsal nuclear gradient¹⁶. Consequently, there is no *sog* expression, and *dpp* is expressed ubiquitously in both dorsal and ventral regions¹⁷. In these mutants, the stripe-2-*sog* transgene provides the only source of Sog products (Fig. 4c; and see later) and these mediate a substantial reorganization in Dpp/Screw signalling activity.

Race was expressed in anterior regions of *gd* mutant embryos (Fig. 2a–c), possibly because of head-specific factors that enhance Dpp signalling. Introduction of the stripe-2-*sog* transgene inhibited *Race* expression within and on either side of the *sog* stripe (Fig. 2d–f). In addition, staining was augmented far from the stripe; this was particularly evident in posterior regions in which *Race* expression is normally lost (Fig. 2e; compare with Fig. 2b). *Race* expression was sustained in advanced-stage embryos (Fig. 2f; compare with Fig. 2c).

The simplest interpretation of these results is that high concentrations of Sog secreted near the stripe inhibit Dpp/Screw signalling, whereas low levels that diffuse far from the stripe augment signalling and activate *Race* expression. However, it is conceivable that the upregulation of *Race* depends on a relay mechanism. The inhibition of Dpp/Screw signalling might trigger formation of the dorsal epidermis in the vicinity of the stripe-2-*sog* pattern. These epidermal cells could then induce neighbouring regions to form amnioserosa through an unknown signal. We investigated this possibility by manipulating the number of copies of the *dpp*⁺ gene in *gd*[−] embryos (Fig. 3). Embryos were stained with a mixture of *sog* and *Race* RNA probes. The stripe-2-*sog* transgene failed to upregulate *Race* expression in embryos carrying just one copy of *dpp* (Fig. 3a; compare with Fig. 3b). There was a general increase in *Race* expression in embryos carrying three copies of *dpp* (Fig. 3d) as compared with two copies (Fig. 2a). Introduction of the stripe-2-*sog* transgene reorganized the *Race* pattern (Fig. 3c), such that a narrow zone of inhibition was formed in the vicinity of the stripe and the posterior domain of enhanced expression was expanded as compared with embryos carrying two copies of *dpp* (Fig. 3b). These results indicate that the specification of cell types is uncoupled by

low versus high levels of Dpp signalling. A narrow domain of low Dpp signalling activity was associated with a broadened region of amnioserosa specification (Fig. 3c), whereas a broad domain of low Dpp signalling failed to induce neighbouring cells to express *Race* (Fig. 3a).

More definitive evidence against a relay model was obtained by fusing DNA encoding the rat integral-membrane protein CD2 in-frame at the 3' terminus of the *sog* coding sequence. The resulting Sog-CD2 fusion protein should be 'tethered' to the plasma membrane and unable to diffuse over long distances. This transgene created a gap in the normal *Race* expression pattern when introduced into wild-type embryos (similar to the pattern in Fig. 1d), indicating that it retains activity (data not shown). The modified *sog* transgene also inhibited the anterior *Race* expression pattern in *gd*[−] embryos (Fig. 4b, arrow; compare with Fig. 4a), but did not augment staining in posterior regions (Fig. 4 legend). In contrast, the unmodified stripe-2-*sog* transgene both inhibited the anterior *Race* pattern (Fig. 4d, arrow) and upregulated posterior *Race* expression (Fig. 4d, arrowhead). These results argue against a relay mechanism and indicate that Sog diffusion may be essential in augmenting Dpp/Screw signalling.

Peak Dpp signalling may depend on the protease Tld⁵. *tld* is expressed in the dorsal ectoderm of wild-type embryos⁴, and is expressed ubiquitously in *gd* mutants. The stripe-2-*sog* transgene was expressed in *gd*[−] embryos heterozygous for a *tld* mutation. Mutant embryos were stained with a mixture of *sog* and *Race* hybridization probes (Fig. 5). *Race* expression was reduced in posterior regions (Fig. 5b) as compared with *gd*[−] embryos containing two wild-type copies of *tld* (Fig. 5a). These results indicate that peak Dpp signalling may depend on the proteolytic degradation of an inhibitory complex containing Sog. In contrast, the stripe-2-*sog* transgene continued to inhibit *Race* expression (Fig. 5b), suggesting that Tld is dispensable for this activity.

To confirm that Sog is specifically required for peak Dpp signalling, we used the *Xenopus* protein Noggin¹⁰ as a heterologous inhibitor to sequester Dpp into an inactive complex. Noggin

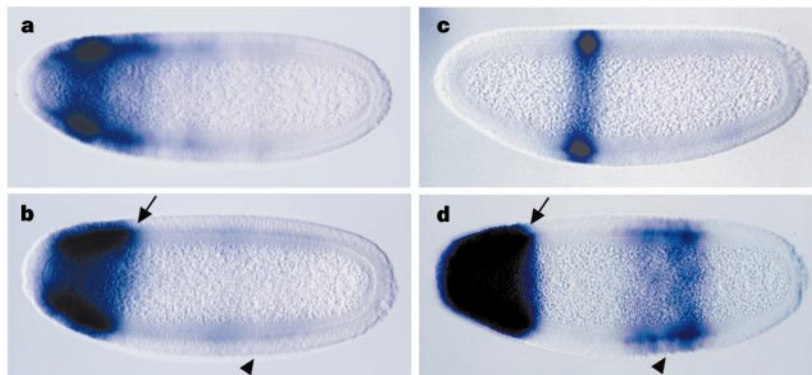


Figure 4 A tethered form of Sog does not mediate long-range enhancement of *Race* expression. The *gd*[−] embryos used carried either the wild-type stripe-2-*sog* transgene (**c, d**) or a modified transgene encoding the integral-membrane protein CD2 fused in-frame to the 3' terminus of the *sog* coding region (**b**). Embryos were stained with either the *Race* (**a, b, d**) or *sog* (**c**) RNA hybridization probe. **a**, Cellularized *gd*[−] embryo stained with the *Race* probe. The most intense staining is detected in anterior regions. **b**, As for **a**, except that the embryo contains the modified stripe-2-*sog*-CD2 transgene. *Race* expression is restricted to anterior regions (arrow), indicating that the transgene may be able to inhibit Dpp/Screw signalling (compare with **a**). However, there is no increase in the *Race* pattern in posterior regions (arrowhead), in contrast to the situation observed with the unmodified, untethered form of the *sog* transgene (**d**). This result indicates that the Sog-CD2 fusion protein may be unable to diffuse and activate *Race*. However, there appears to be some release of Sog from the modified *sog*-CD2 fusion gene, as about half of all older embryos exhibit variable increases in the posterior *Race* pattern. Nonetheless, the wild-type stripe-2-*sog* transgene never produces this

pattern in which the anterior domain is repressed and there is no enhancement of expression (**d**). Perhaps the heterologous Sog-CD2 fusion protein is subject to proteolytic cleavage during development. Although unlikely, we cannot completely exclude the possibility that low levels of Sog result in an intermediate threshold of Dpp/Screw which specifies dorsal epidermis and triggers the expression of signalling factors that specify neighbouring cells to form amnioserosa. **c**, A cellularized *gd*[−] embryo that contains the unmodified stripe-2-*sog* transgene. The embryo was stained with the *sog* RNA probe to show that the transgene is the sole source of *sog* expression in these mutants. Some embryos exhibit weak and variable staining within the limits of stripe 7 (not shown). In addition, the stripe-2-*sog* expression pattern is broader in younger embryos, but quickly refines into the sharp stripe seen here (data not shown). **d**, *Race* staining pattern in a cellularized *gd*[−] embryo that contains the unmodified stripe-2-*sog* transgene. As shown in Fig. 2d, *Race* is inhibited in a broad band in the vicinity of the stripe, but is upregulated in both anterior (arrow) and posterior (arrowhead) regions far from the stripe.

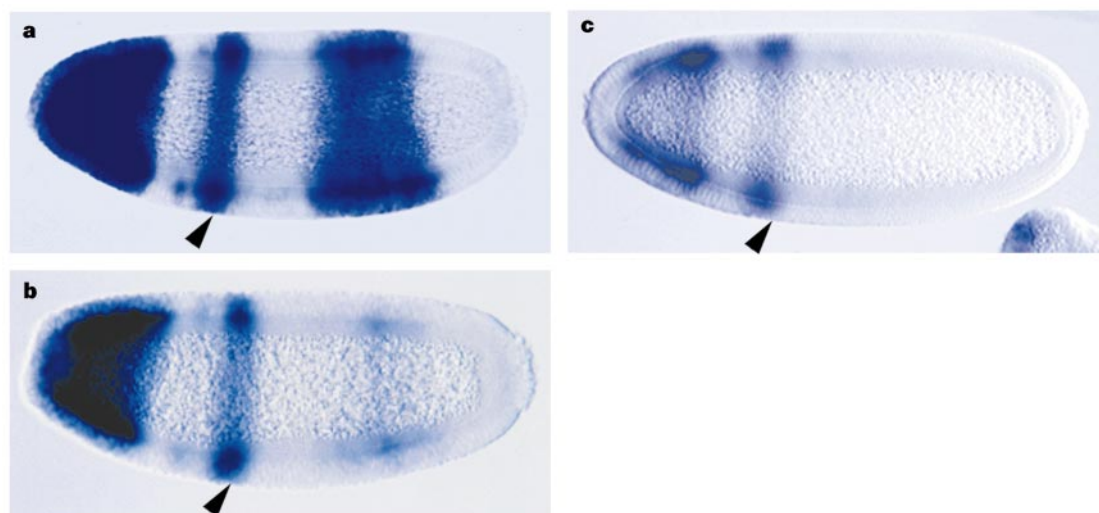


Figure 5 Uncoupling *sog*-mediated enhancement and inhibition of Dpp signalling. The cellularizing *gd*⁺ embryos used contained either a stripe-2-*sog* or stripe-2-*noggin* transgene, and were stained with a mixture of *Race* and *sog* (a, b) or *Race* and *noggin* (c) hybridization probes. The arrowheads indicate the stripe-2 pattern. a, The embryo contains the *sog* transgene. As shown in Fig. 3b, the stripe-2 *sog* pattern is asymmetrically positioned between the anterior and posterior *Race* expression domains. b, As for a, except that the embryo is heterozygous for a moderate *tld* mutation²⁶ (*tld*^{9B}/+). The stripe-2-*sog* transgene

continues to inhibit *Race* expression, but there is a substantial loss of staining in posterior regions (compare with a). c, The embryo contains the stripe-2-*noggin* transgene. *Race* expression is restricted to anterior regions, indicating that Noggin can inhibit, but not enhance, Dpp/Screw signalling. The stripe-2-*noggin* transgene is poorly expressed; however, whenever a *noggin* stripe was observed, there was usually some indication of *Race* inhibition. The embryo shown in this panel represents one of the best examples of robust *noggin* expression.

binds Dpp (and possibly Screw) but is not recognized by Tld, thereby preventing the release of active Dpp^{18,19}. Noggin was expressed in the *Drosophila* embryo using the same *eve* stripe-2 FLP-FRT system⁹ that we used to misexpress *sog*. *Race* and *noggin* expression were visualized by *in situ* hybridization. Embryos bearing the transgene expressed a stripe of *noggin* that had the same inhibitory effect on *Race* expression as stripe-2-*sog* (Fig. 5c). However, there was no upregulation of *Race* in posterior regions. These results show that Noggin can mimic the inhibitory effect of Sog, but is unable to generate the peak Dpp signalling output.

We have shown that a localized stripe of *sog* expression is sufficient to reorganize the dorsoventral pattern of mutant embryos lacking a Dorsal nuclear gradient. Cells located near the Sog source exhibit repression of dorsal-ectoderm marker genes, whereas those located far from the source (>10 cell diameters) exhibit peak Dpp signalling, including upregulation of *Race* expression and the restoration of amnioserosa in advanced-stage embryos. Different models have been proposed to explain the requirement for Sog in generating peak Dpp activity. One invokes the diffusion of Sog–Dpp or Sog–Screw complexes away from the ventrolateral Sog stripes, thereby focusing Dpp and/or Screw at the dorsal midline^{6,7,19}. An alternative model is that a product resulting from the cleavage of Sog directly signals formation of the amnioserosa⁵, possibly by augmenting the binding of Dpp or Screw to the receptors Thick veins and Saxophone.

Results from studies of *Xenopus* and *Drosophila* indicate that extracellular signalling molecules of the TGF- β family can generate different thresholds of gene activity through a classical ‘French flag’ mechanism of positional information²⁰. Cells located near the source of signal exhibit a peak threshold of gene activity, whereas those located progressively farther from the source express target genes that can be activated only by low levels of signal. A localized source of activin leads to different patterns of *gooseoid* and *Xbra* expression in *Xenopus* animal caps^{21,22}, and a localized source of Dpp at the boundary between the anterior and posterior compartments of *Drosophila* wing imaginal discs leads to differential patterns of *spalt* and *optomotor-blind* expression^{23,24}. Our results indicate that the patterning of the *Drosophila* dorsal ectoderm may depend on a

different mechanism^{5,19}. In the absence of Sog, Dpp signalling is just below the critical threshold required for the specification of the amnioserosa. The activation of amnioserosa-specific genes, such as *Race*, depends on enhancing Dpp/Screw signalling. This can be achieved by overexpressing Dpp²⁵ (Fig. 3d), but is normally achieved with a localized source of Sog. Given the evolutionary conservation of Sog and Dpp (chordin and bone morphogenetic protein in *Xenopus*)²⁶, it is conceivable that this mechanism is used generally in metazoans. □

Methods

Plasmid construction, P-element transformation and *in situ* hybridization. A HindIII–NotI fragment of the *sog* cDNA (nucleotides 0–4,541) from pBSsog (a gift from E. Bier) and an EcoRI–EcoRV *noggin* cDNA fragment from pNoggin Δ 5 (ref. 10) were blunt-end-ligated into 22FPE (ref. 9) (provided by S. Small). Rat *CD2* cDNA was fused to *sog* in pBSsog by first mutagenizing the *sog* stop codon by the polymerase chain reaction (PCR), creating a unique Asp718 site. A PstI fragment of rat *CD2* cDNA (from FC15; a gift from G. Struhl) lacking the signal sequence was inserted in-frame into the Asp718 site in a blunt-ended ligation. The *sog*–*CD2* cDNA was then transferred into 22FPE. P-element-mediated transformation by injection into *yw*^{67c23} embryos was as described²⁷. Multiple lines were tested for all constructs. *In situ* hybridizations using digoxigenin-labelled antisense RNA probes and alkaline phosphatase substrate were performed as described^{11,28}.

***Drosophila* stocks and genetic crosses.** The fly stocks used were as follows: a transgenic line homozygous for a P[ry⁺ β -tubulin–*flp*] insertion provided by G. Struhl); *gastrulation defective*, *gd*⁷/FM3; *decapentaplegic*, *dpp*^{Hin37}/GlaDp(2;2)DTD48; *short gastrulation*, *sog*⁵⁶/FM7c; and *tolloid*, *tld*^{9B}/TM3. Females containing the misexpression constructs were crossed to males carrying the β -tubulin–*flp* gene to obtain males containing both transgenes; in these males the misexpression constructs are activated by the spermatocyte-specific removal of a flip-out cassette catalysed by β -tubulin–*flp*. These males were crossed to *yw*^{67c23}, *gd*⁷/gd⁷, *sog*⁵⁶/FM7c (Fig. 1f), *gd*⁷/gd⁷; GlaDp(2;2)DTD48/+ (Fig. 3c) or *gd*⁷/gd⁷; *dpp*^{Hin37}/GlaDp(2;2)DTD48 (Fig. 3a) females and embryos were collected and analysed by *in situ* hybridization. *gd*⁷/gd⁷; GlaDp(2;2)DTD48/+ females, which contain an extra copy of *dpp*, were obtained by first crossing *dpp*^{Hin37}/GlaDp(2;2)DTD48 males with *gd*⁷/FM3 females. *gd*⁷; GlaDp(2;2)DTD48/+ males from this cross were

then mated with *gd⁷/FM3* females to obtain *gd⁷/gd⁷*; *GlaDp(2;2)DTD48/+* females. This cross also generated *gd⁷/FM3*; *GlaDp(2;2)DTD48/+* females, which were first crossed to *dpp^{Hin37}/GlaDp(2;2)DTD48* males. The resulting males (including *gd⁷*; *dpp^{Hin37}/GlaDp(2;2)DTD48* males) were backcrossed to *gd⁷/FM3*; *GlaDp(2;2)DTD48/+* females. From this cross, females that lacked the FM3 balancer (including *gd⁷/gd⁷*; *dpp^{Hin37}/GlaDp(2;2)DTD48* females) were crossed to *P[st2-sog]*; *P[ry⁺ β ₂-tubulin-*flp*]* males to obtain *st2-sog*, *gd⁷*, *dpp/+* embryos.

The dose of *tld* was lowered by first crossing *P[st2-sog]* females to *tld^{9B}/TM3* males. *P[st2-sog]*; *tld^{9B}* females were then crossed to *P[ry⁺ β ₂-tubulin-*flp*]* males generate *P[st2-sog]*; *tld^{9B}/P[ry⁺ β ₂-tubulin-*flp*]* males, which can activate the transgene. Embryos were collected from crosses between these males and *gd⁷/gd⁷*; *tld^{9B}* females, obtained by crossing *gd⁷/FM3* females with *tld^{9B}/TM3* males then backcrossing *gd⁷*; *tld^{9B}* males with *gd⁷/FM3* females. The embryos shown in Figs 3, 5 were represented at the predicted frequency based on the crosses.

Received 23 October 1998; accepted 8 February 1999.

- Ferguson, E. L. & Anderson, K. V. *decapentaplegic* acts as a morphogen to organize dorsal-ventral pattern in the *Drosophila* embryo. *Cell* **71**, 451–461 (1992).
- Wharton, K. A., Ray, R. P. & Gelbart, W. M. An activity gradient of *decapentaplegic* is necessary for the specification of dorsal pattern elements in the *Drosophila* embryo. *Development* **117**, 807–822 (1993).
- François, V., Solloway, M., O'Neill, J. W., Emery, J. & Bier, E. Dorsal-ventral patterning of the *Drosophila* embryo depends on a putative negative growth factor encoded by the *short gastrulation* gene. *Genes Dev.* **8**, 2602–2616 (1994).
- Shimell, M. J., Ferguson, E. L., Childs, S. R. & O'Connor, M. B. The *Drosophila* dorsal-ventral patterning gene *tolloid* is related to human bone morphogenetic protein 1. *Cell* **67**, 469–481 (1991).
- Marques, G. *et al.* Production of a Dpp activity gradient in the early *Drosophila* embryo through the opposing actions of the Sog and Tld proteins. *Cell* **91**, 417–426 (1997).
- Neul, J. L. & Ferguson, E. L. Spatially restricted activation of the Sax receptor by Scw modulates Dpp/Tkv signalling in *Drosophila* dorsal-ventral patterning. *Cell* **95**, 483–494 (1998).
- Nguyen, M., Park, S., Marques, G. & Arora, K. Interpretation of a BMP activity gradient in *Drosophila* embryos depends on synergistic signalling by two type I receptors, Sax and Tkv. *Cell* **95**, 495–506 (1998).
- Zusman, S. B., Sweeton, D. & Wieschaus, E. F. *short gastrulation*, a mutation causing delays in stage specific cell shape changes during gastrulation in *Drosophila melanogaster*. *Dev. Biol.* **129**, 417–427 (1988).
- Kosman, D. & Small, S. Concentration-dependent patterning by an ectopic expression domain of the *Drosophila* gap gene *knirps*. *Development* **124**, 1343–1354 (1997).
- Smith, W. C. & Harland, R. M. Expression cloning of noggin, a new dorsalising factor localized to the Spemann organizer in *Xenopus* embryos. *Cell* **70**, 829–840 (1992).
- Jiang, J., Kosman, D., Ip, Y. T. & Levine, M. The dorsal morphogen gradient regulates the mesoderm determinant *twist* in early *Drosophila* embryos. *Genes Dev.* **5**, 1881–1891 (1991).
- Tatei, K., Cai, H., Ip, Y. T. & Levine, M. Race: a *Drosophila* homologue of the angiotensin converting enzyme. *Mech. Dev.* **51**, 157–168 (1995).
- Rusch, J. & Levine, M. Regulation of a *dpp* target gene in the *Drosophila* embryo. *Development* **124**, 303–311 (1997).
- Arora, K., Levine, M. S. & O'Connor, M. B. The *screw* gene encodes a ubiquitously expressed member of the TGF- β family required for specification of dorsal cell fates in the *Drosophila* embryo. *Genes Dev.* **8**, 2588–2601 (1994).
- Konrad, K. D., Goralski, T. J., Mahowald, A. P. & Marsh, J. L. The *gastrulation-defective* gene of *Drosophila melanogaster* is a member of the serine protease superfamily. *Proc. Natl Acad. Sci. USA* **95**, 6819–6824 (1998).
- Roth, S., Stein, D. & Nusslein-Volhard, C. A gradient of nuclear localization of the dorsal protein determines dorsoventral pattern in the *Drosophila* embryo. *Cell* **59**, 1189–1202 (1989).
- Rusch, J. & Levine, M. Threshold responses to the dorsal regulatory gradient and the subdivision of primary tissue territories in the *Drosophila* embryo. *Curr. Opin. Genet. Dev.* **6**, 416–423 (1996).
- Piccolo, S. *et al.* Cleavage of chordin by xolloid metalloprotease suggests a role for proteolytic processing in the regulation of Spemann organizer activity. *Cell* **91**, 407–416 (1997).
- Holley, S. A. *et al.* The *Xenopus* dorsalizing factor noggin ventralises *Drosophila* embryos by preventing Dpp from activating its receptor. *Cell* **86**, 607–617 (1996).
- Lewis, J., Slack, J. M. & Wolpert, L. Thresholds in development. *J. Theor. Biol.* **65**, 579–590 (1977).
- McDowell, N., Zorn, A. M., Crease, D. J. & Gurdon, J. B. Activin has direct long-range signalling activity and can form a concentration gradient by diffusion. *Curr. Biol.* **7**, 671–681 (1997).
- Jones, C. M., Armes, N. & Smith, J. C. Signalling by TGF- β family members: short-range effects of Xnr-2 and BMP-4 contrast with the long-range effects of activin. *Curr. Biol.* **6**, 1468–1475 (1996).
- Nellen, D., Burke, R., Struhl, G. & Basler, K. Direct and long-range action of a DPP morphogen gradient. *Cell* **85**, 357–368 (1996).
- Lecuit, T. *et al.* Two distinct mechanisms for long-range patterning by Decapentaplegic in the *Drosophila* wing. *Nature* **381**, 387–393 (1996).
- Ferguson, E. L. & Anderson, K. V. Localized enhancement and repression of the activity of the TGF- β family member, decapentaplegic, is necessary for dorsal-ventral pattern formation in the *Drosophila* embryo. *Development* **114**, 583–597 (1992).
- Holley, S. A. *et al.* A conserved system for dorsal-ventral patterning in insects and vertebrates involving Sog and chordin. *Nature* **376**, 249–253 (1995).
- Rubin, G. & Spradling, A. Genetic transformation of *Drosophila* with transposable element vectors. *Science* **218**, 348–353 (1982).
- Tautz, D. & Pfeifle, C. A non-radioactive *in situ* hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals a translational control of the segmentation gene hunchback. *Chromosoma* **98**, 81–85 (1989).

Acknowledgements. We thank R. Harland for the *noggin* cDNA and suggesting the stripe-2–*noggin* experiment; and G. Struhl, R. Harland, Mark Ashe and members of the Levine laboratory for helpful discussions and encouragement. This work was funded by a grant from the NIH.

Correspondence and requests for materials should be addressed to M.L. (e-mail: mlevine@uclink4.berkeley.edu).

A new secreted protein that binds to Wnt proteins and inhibits their activities

Jen-Chih Hsieh*†‡, Laurent Kodjabachian§, Martha L. Rebbert§, Amir Rattner*†, Philip M. Smallwood*†, Cynthia Harryman Samos||, Roel Nussell, Igor B. Dawid§ & Jeremy Nathans*†‡

* Department of Molecular Biology and Genetics, † Howard Hughes Medical Institute, and ‡ Departments of Neuroscience and Ophthalmology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

§ Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, USA

|| Department of Developmental Biology, Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94035, USA

‡ These authors contributed equally to this work.

The Wnt proteins constitute a large family of extracellular signalling molecules that are found throughout the animal kingdom and are important for a wide variety of normal and pathological developmental processes^{1,2}. Here we describe Wnt-inhibitory factor-1 (WIF-1), a secreted protein that binds to Wnt proteins and inhibits their activities. WIF-1 is present in fish, amphibia and mammals, and is expressed during *Xenopus* and zebrafish development in a complex pattern that includes paraxial presomitic mesoderm, notochord, branchial arches and neural crest derivatives. We use *Xenopus* embryos to show that WIF-1 overexpression affects somitogenesis (the generation of trunk mesoderm segments), in agreement with its normal expression in paraxial mesoderm. *In vitro*, WIF-1 binds to *Drosophila* Wingless and *Xenopus* Wnt8 produced by *Drosophila* S2 cells. Together with earlier results obtained with the secreted Frizzled-related proteins^{1,2}, our results indicate that Wnt proteins interact with structurally diverse extracellular inhibitors, presumably to fine-tune the spatial and temporal patterns of Wnt activity.

There are two families of secreted molecules known to inhibit Wnt signalling activity: the secreted Frizzled-related protein (sFRP) family, whose members all have an amino-terminal cysteine-rich domain (CRD) that is highly homologous to the ligand-binding domain of Frizzled proteins, which are transmembrane Wnt receptors^{1,2}; and the protein family Dickkopf (Dkk), whose mechanism of action is at present unknown³.

WIF-1 was first identified as an expressed sequence tag from the human retina (J. P. Macke, P.M.S. and J.N., unpublished results), and highly conserved orthologues have been isolated from mouse, *Xenopus* and zebrafish (Fig. 1). The deduced amino-acid sequence of WIF-1 has an N-terminal signal sequence, a domain of ~150 amino acids (the WIF domain), five epidermal growth factor (EGF)-like repeats that are most similar to those of the extracellular matrix protein tenascin, and a hydrophilic domain of ~45 amino acids at the carboxy terminus.

In the adult mouse, WIF-1 expression is highest in the heart and lung, and lower in the brain and eye (Fig. 2a). Northern-blot hybridization with *Xenopus* total RNA revealed the presence of a single transcript expressed first at neurula stages (not shown), and *in situ* hybridization to *Xenopus* or zebrafish embryos confirmed that no messenger RNA is detectable at the gastrula stage. In *Xenopus*, WIF-1 is expressed during somitogenesis predominantly in the unsegmented paraxial presomitic mesoderm and to a much lesser extent in newly segmented somites (Fig. 2b–e). In zebrafish, WIF-1 is highly expressed in unsegmented paraxial mesoderm and is virtually undetectable in mature somites (Fig. 2g–i). WIF-1 expression is visible in both species in the notochord in register

with mature somites, but not with unsegmented paraxial mesoderm (Fig. 2d–i). WIF-1 mRNA is also present in discrete domains of the anterior brain (Fig. 2f, j), as well as in the visceral arches, nasal placodes and otic vesicles (Fig. 2e, f).

To investigate the activity of WIF-1 *in vivo*, we injected early *Xenopus* embryos with synthetic RNA encoding human WIF-1 (hWIF-1) either alone or in combination with developmental regulators. Overexpression of hWIF-1 provoked anteriorization and dorsalization of the embryos, as revealed by the formation of secondary body axes upon ventral injection and by the enlargement of antero–dorsal structures upon dorsal injection (Fig. 3a–f). Such phenotypes are also seen in embryos in which the late ventralizing XWnt8 pathway has been blocked by a dominant-negative Wnt ligand or by overexpression of the secreted Wnt inhibitor sFRP-3/ Frzb (refs 1,2,4). By using Wnt-dependent axis induction in early *Xenopus* embryos as an assay, we found that hWIF-1 blocks the activity of XWnt8 in a dose-dependent manner (Table 1). In this assay, the N-terminal WIF domain (WD) of hWIF-1 is as effective as full-length hWIF-1. Inhibition also occurs following injection of XWnt8 and hWIF-1 separately in adjacent blastomeres, suggesting that the interaction occurs in the extracellular space (Table 1). In a second measure of XWnt8 inhibition, the late ventralizing activity of XWnt8, generated by overexpression from DNA expression constructs at gastrulation^{5,6}, was also antagonized by co-injection of DNA encoding hWIF-1 (not shown). The inhibition is specific to Wnt ligands because axis induction by the transcription factor

Siamois, which is activated by the Wnt pathway, is not inhibited, nor is the secreted dorsalizing agent Chordin (Table 1).

hWIF-1 inhibits XWnt8 more than Wingless (Wg), and XWnt3A is only weakly inhibited (Table 1). These results place WIF-1 in the same category as sFRP-3/Frzb, as a putative secreted Wnt antagonist. However, the activities of these two factors differ when measured by co-injection with the BMP inhibitor Chordin. Co-injection of Chordin with sFRP-3/Frzb promoted complete secondary axis formation at low frequency and ectopic heads were always cyclopic, as reported earlier³ (Fig. 3g; Table 1), whereas co-injection of Chordin and hWIF-1 gave complete secondary axes in all cases, with most having two eyes (Fig. 3h; Table 1). This difference is probably qualitative rather than quantitative as we were unable to obtain two-eyed secondary heads with Chordin and higher doses of sFRP-3/Frzb. The WIF domain alone could also synergize with Chordin to induce ectopic heads, but it did not promote separation of the eyes (Fig. 3i; Table 1), suggesting that the EGF-like repeats are necessary for full activity of WIF-1.

As WIF-1 is expressed in paraxial mesoderm in both zebrafish and *Xenopus*, we tested the effect of WIF-1 overexpression on somite formation by injecting one side of *Xenopus* 4-cell embryos with hWIF-1 DNA to initiate expression by the onset of gastrulation: we found that hWIF-1 specifically antagonized the formation of the most posterior presomitic mesoderm (Fig. 3j, k). At later stages of development, hWIF-1-injected embryos consistently had abnormally shaped and disorganized somites in which the process of

hWIF-1	MARRSAFPAA	ALWLWSILLC	LLALRAEAGP	PQEESELYLWI	DAHQARVLIG	50
mWIF-1	---R---F	--R---P-	--L---D-Q	-P-----	-----	50
xWIF-1	M-LTGYF	-AP-C--F-F	I--.H-D-Q	.-D--M--	-----	45
zWIF-1	MAFRT--V	Q-H-KACV-L	--GGLL--AY	EERGTM-M--	--N---I--	48
hWIF-1	FEEDILIVSE	GKMAPFTHDF	RKAQQRMPAI	PVNIHSMNFT	WQAAGQAEYF	100
mWIF-1	-----	-----	-----	-----	-----	100
xWIF-1	-----A-	-----	-----	-----A---	---T-----	95
zWIF-1	-----	-----	-----	-----HV--	---TD-----	98
hWIF-1	YEFLSLRSLD	KGIMADPTVN	VPLLGTVPKH	ASVVQVGFPC	LGKQDGVAAF	150
mWIF-1	-----	-----	-----	-----	-----	150
xWIF-1	-----	-----	M-----	-T-I-----	--N-----	145
zWIF-1	---QT-----	-D--D-----	---S-----	-----	R-D-----	148
hWIF-1	EVDVIVMNSE	GNTILQTPQN	AIFFKTCLQA	ECPGGCRNGG	FCNERRICEC	200
mWIF-1	--N-----	-----R---	-----Q---	-----	-----V---	200
xWIF-1	--N-----	--V-----	-----Q---	K-T-----	---D-HV---	195
zWIF-1	--TIL--DAG	--I--R--H-	-----QR-	K-P-----	Y----QV---	198
hWIF-1	PDGFHGHPC	KALCTPRCMN	GGLCVTPGFC	ICPPGFYGVN	CDKANCSTTC	250
mWIF-1	---Y-----	---I-----	-----	-----	-----	250
xWIF-1	---Y-----	---M-----	-----L-	---Y--I-	---V--T-H-	245
zWIF-1	Q---Y-V---	---S---L-	---MS--V-	---YF-SS	-ER-----	248
hWIF-1	FNGGTCFYPG	KCICPPGLEG	EQCEISKCPQ	PCRNGGKCIG	KSKCKCSKGY	300
mWIF-1	-----	-----	-----L---	-----	-----P---	300
xWIF-1	L-----	-----S-Y--	---T---Q-	-----S-	-N-----	295
zWIF-1	L-----H--	---AVSF--	VR--L---R-	-----T-	RN-----	298
hWIF-1	QGDLCSKPVC	EPGCGAHGTC	HEPNKCQCQE	GWHGRHCNKR	YEASLIHALR	350
mWIF-1	-----	-----	-----R-	-----	-G--M--P-	350
xWIF-1	-----	--S-----	I-----K-	--N--Y--K	-GSN-MN---	345
zWIF-1	H-----A--	--S-----	V---R---R-	-----	FRGGVSNSQ-	348
hWIF-1	PAGALRLQ.-	TPSLKKAEEER	RDPPESENYIW			379
mWIF-1	---G-ER.-	-----D-	-----			379
xWIF-1	-T-SRN--.-	---P-RT-D-	QAL-----			374
zWIF-1	VSPSKHKSPS	VAAA-E-P-T	SQ-S-T--VV			378

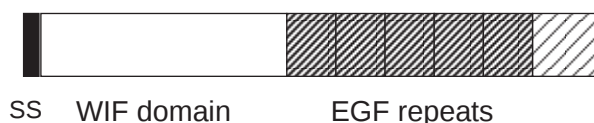


Figure 1 Sequence alignment and structure of vertebrate WIF-1. Top, alignment of the WIF-1 proteins from human (h), mouse (m), *Xenopus* (x), and zebrafish (z). Dashes indicate amino-acid identity and dots indicate a gap. The locations of the

five epidermal growth factor (EGF) repeats are indicated by lines above the sequence. Bottom, domain structure of WIF-1, showing, from N to C terminus: SS, signal sequence; WIF domain; five EGF repeats; and a hydrophilic C terminus.

somite rotation was impaired on the injected side and was occasionally accompanied by fusions of adjacent somites (Fig. 3l–p). We suggest that WIF-1 is involved in segmentation of the mesoderm, perhaps by antagonizing an as-yet unidentified Wnt molecule.

To assess the inhibitory effect of WIF-1 on Wnt signalling in a

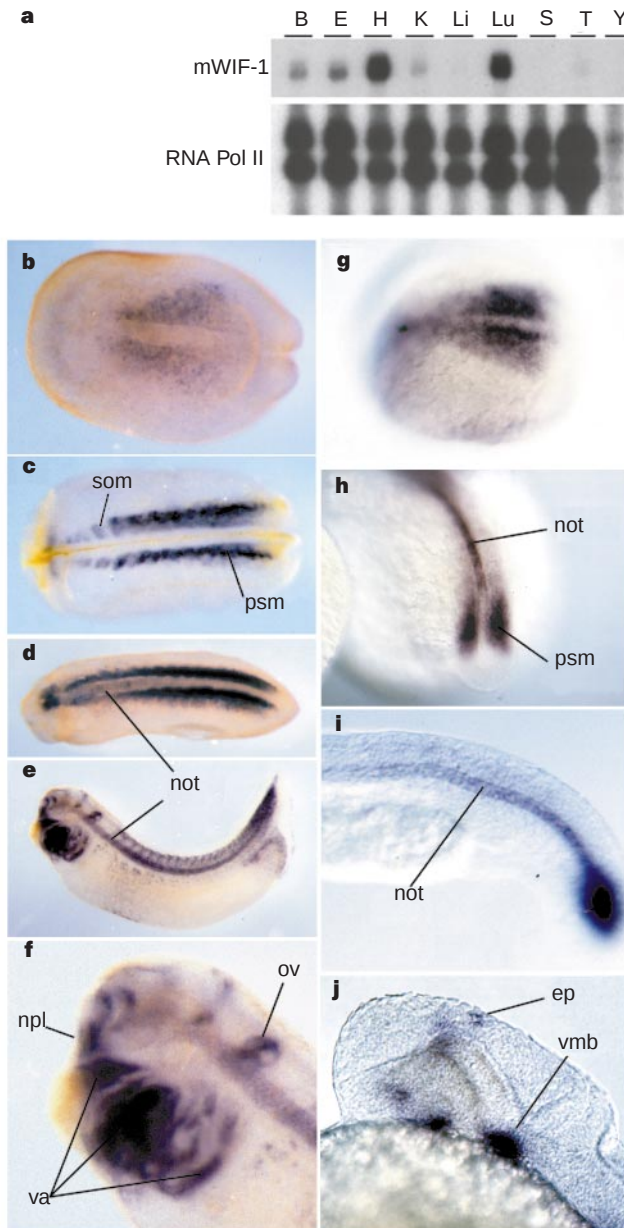


Figure 2 Expression of WIF-1 in the adult mouse and during *Xenopus* and zebrafish development. **a**, RNase protection assay using a mWIF-1 probe (top) or a mouse RNA polymerase II large-subunit probe (bottom); B, brain; E, eye; H, heart; K, kidney; Li, liver; Lu, lung; S, spleen; T, testis; Y, yeast tRNA. **b–j**, WIF-1 mRNAs detected by *in situ* hybridization. **b–f**, *Xenopus*; **g–j**, zebrafish. Expression is first detectable at the onset of somitogenesis in paraxial mesoderm (**b**, **g**). Expression is prominent in unsegmented presomitic mesoderm (psm), and is much weaker in newly formed somites (som; **c**, **d**, **h**, **i**). WIF-1 is expressed in the notochord (not) of 15-somite-stage zebrafish embryos (**h**) and stage-24 *Xenopus* embryos (**d**) in register with mature somites; no expression is detected in notochord regions flanked by unsegmented paraxial mesoderm. This pattern persists until the late stages of somitogenesis (**e**, **i**). In *Xenopus* tadpoles, strong WIF-1 expression is seen in visceral arches (va), and also in otic vesicles (ov), nasal placodes (npl) and several regions of the anterior brain (**f**). In the 24-h zebrafish embryo, expression is seen in the epiphysis (ep) and ventral midbrain (vmb) (**j**).

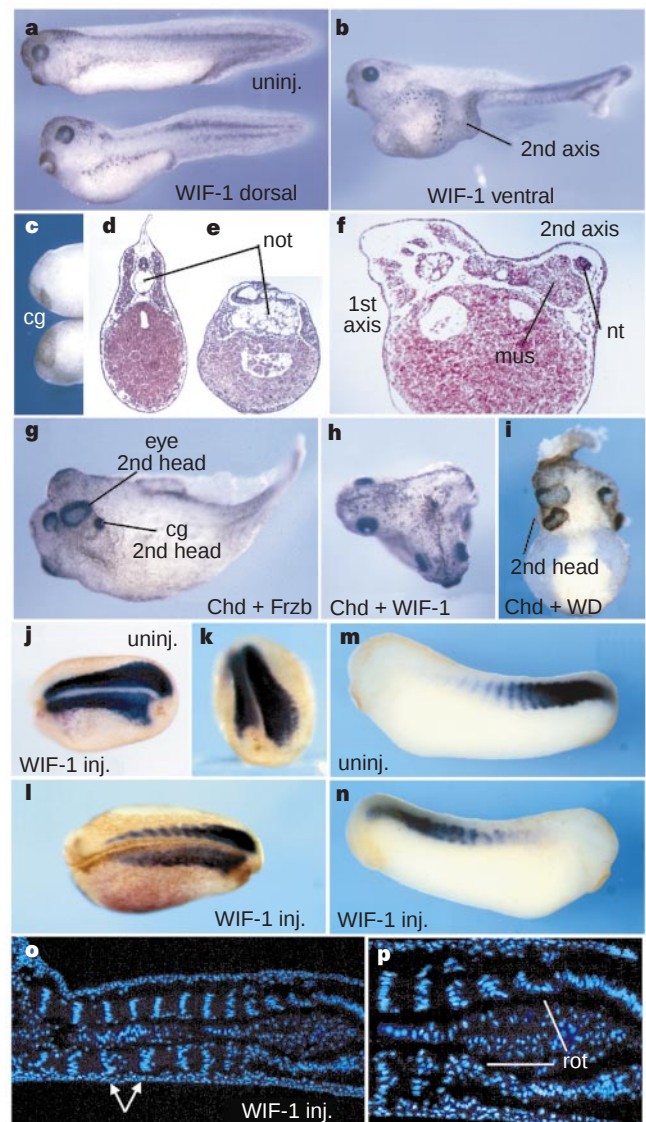


Figure 3 WIF-1 inhibits Wnt signaling and affects somite formation in *Xenopus* embryos. **a–f**, WIF-1 dorsalizes *Xenopus* embryos. hWIF-1 RNA (500 pg per blastomere) was injected into dorsal blastomeres (**a**, **c**, **e**) or ventral blastomeres (**b**, **f**) at the 4-cell stage. Dorsal injections led to anteriorization and hyperdorsalization of the embryos, as revealed by the formation of an enlarged head at the tadpole stage (**a**), an expanded cement gland (cg) field at the neurula stage (**c**), and an enlarged notochord visible in sections (**e**; control embryo in **d**). Ventral injections led to the formation of a partial secondary axis in 20% of the embryos (**b**) containing muscle (mus) and neural tissue (nt) (**f**). **g–i**, WIF-1 synergizes with chordin to generate a complete secondary axis (Table 1). **g**, Co-injection of chordin and *Xenopus* sFRP-3/Frzb RNAs gave rise to embryos with a complete but cyclopic secondary axis. **h**, By contrast, co-injection of chordin and hWIF-1 gave rise to embryos with a complete secondary axis; two separated eyes are evident. **i**, Co-injection of chordin and WD RNA does not promote two-eyed ectopic heads. **j–p**, WIF-1 affects the process of somite formation. 4-cell embryos were injected on one side with 100 pg pCS2+/hWIF-1 DNA and β -galactosidase RNA. β -Galactosidase activity is seen as brown dots and embryos were probed for XMyoD (**j–n**). The posterior domain of XMyoD expression in presomitic mesoderm was suppressed by hWIF-1 injection at the late neurula stage (**j**, dorsal view; **k**, posterior view). Segmentation is impaired by hWIF-1 injection, as revealed by XMyoD expression in mature somites of a stage-20 embryo (**l**) and a stage-26 embryo (**m**, **n**), and by a Hoechst-stained horizontal section of a stage-26 embryo (**o**, **p**). Note the disorganization of somites with occasional fusions (arrows in **o**). The process of somite rotation (rot) is delayed and myotome nuclei do not align properly (**p**).

simplified system, we determined the effect of hWIF-1, an IgG heavy-chain fusion containing full-length hWIF-1 (WIF-1-IgG), or the IgG heavy chain alone (IgG) on Wg-dependent Armadillo stabilization in *Drosophila* clone-8 cells (ref. 7; Fig. 4a, b). At 30 nM, WIF-1-IgG or hWIF-1 significantly inhibit the response of clone-8 cells to Wg-conditioned medium. To determine the mechanism of WIF-1 inhibition, we tested whether WIF-1 could

bind to soluble, secreted Wnt proteins. To produce a functional secreted vertebrate Wnt, we transfected S2 cells with a complementary DNA encoding XWnt8 tagged at its C terminus with a Myc epitope (XWnt8-Myc). The resulting conditioned medium contained soluble XWnt8-Myc, which stabilizes Armadillo in *Drosophila* clone-8 cells and β -catenin in C57MG cells, and can bind to the surface of 293 cells transfected with any one of several

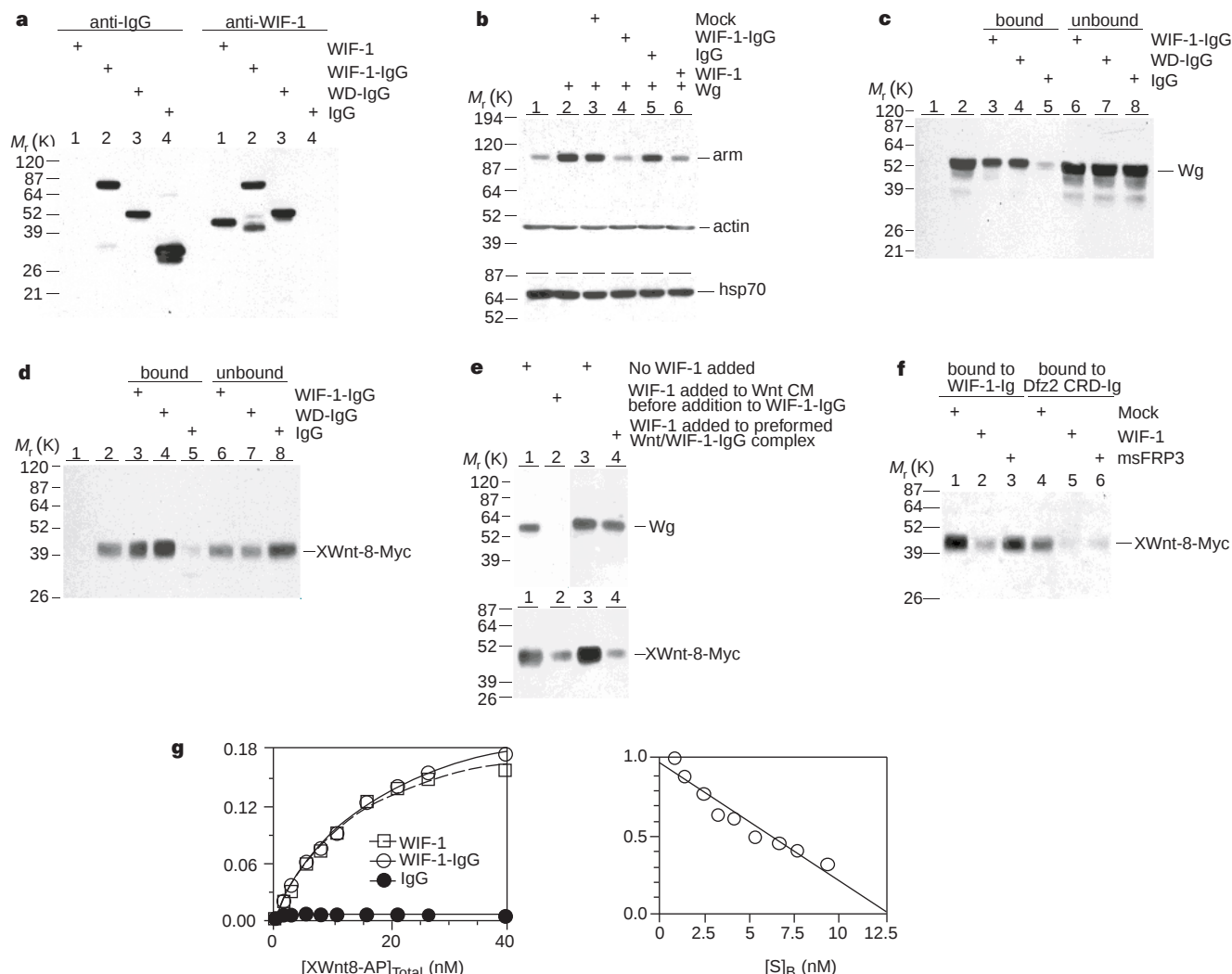


Figure 4 hWIF-1 binds reversibly to soluble Wg and XWnt8-Myc and inhibits the Wg response of clone-8 cells. **a**, Characterization of hWIF-1 and WIF-1-IgG fusion proteins secreted from transiently transfected 293 cells by immunoblotting with anti-human IgG (left) or affinity-purified anti-WIF-domain antibodies (right). Each lane contains 40 μ l 20-fold concentrated serum-free conditioned medium. **b**, hWIF-1 inhibits Wg signalling by clone-8 cells. Clone-8 cells were treated with control (lane 1) or Wg-containing (lanes 2–6) medium from S2 cells that had been preincubated with DMEM/F-12 (lanes 1, 2) or with 0.4 ml control medium (lane 3) or medium containing 30 nM WIF-1-IgG (lane 4), IgG (lane 5), or WIF-1 (lane 6). Similar results were obtained in 6 independent experiments. **c, d**, WIF-1-IgG and WD-IgG bind Wg (**c**) and XWnt8-Myc (**d**). Lane 1, 40 μ l control S2 medium; lane 2, 10% of the Wg or XWnt8-Myc medium used in each binding reaction; lanes 3–5, 30% (**c**) or 40% (**d**) of the bound material; lanes 6–8, 10% of the unbound fraction from the samples in lanes 3–5, respectively. **e**, Reversible binding of WIF-1-IgG to Wg (top) and XWnt8-Myc (bottom). In lanes 1 and 2, 10 pmol WIF-1-IgG was prebound to protein A-Sepharose, and then added to 250 μ l Wg (top) or 200 μ l XWnt8-Myc (bottom) medium that had been preincubated for 2 h at 4 °C with control medium (lane 1) or medium containing 200 pmol hWIF1. After 2 h incubation at 4 °C, the beads were processed as described in Methods. In lanes 3 and 4, 10 pmol WIF-1-IgG prebound to protein A-Sepharose was preincubated with 250 μ l Wg (top) or 200 μ l XWnt8-Myc (bottom) medium for

2 h at 4 °C, and the unbound proteins removed with five washes of PBS. Control medium (lane 3) or medium containing 200 pmol hWIF-1 was added to the beads, which were then incubated for 2 h at 4 °C and washed as for lanes 1 and 2. **f**, hWIF-1 competes efficiently with both WIF-1-IgG and Dfz2CRD-IgG for XWnt8-Myc binding, but msFRP-3 competes efficiently only with Dfz2CRD. 100 μ l XWnt8-Myc medium was incubated at 4 °C for 1 h with 100 μ l control medium (lanes 1, 4), 100 μ l medium containing 200 pmol hWIF-1 (lanes 2–5) and 200 pmol purified SFRP-3 (lanes 3–6), 5 pmol WIF-1-IgG (lanes 1–3) or Dfz2CRD-IgG (lanes 4–6) were incubated with protein A-Sepharose, and the washed beads were then added to each sample for an additional 2 h incubation at 4 °C. The XWnt8-Myc/WIF-1-IgG complex is resistant, whereas the XWnt8-Myc/Dfz2CRD-IgG complex is susceptible to msFRP-3 competition. Both complexes are susceptible to competition by hWIF-1. **g**, Left, binding of XWnt8-AP to WIF-1 (squares, dashed line), WIF-1-IgG (circles, solid line) and IgG (filled circles, solid line). The vertical axis shows the increase in absorbance at 405 nm following cleavage of *p*-nitrophenylphosphate. Right, Scatchard plot of the data for XWnt8-AP binding to WIF-1-IgG. The data for XWnt8-AP binding to WIF-1 give an almost identical best-fit line (not shown). Vertical axis, ratio of bound to free XWnt8-AP; horizontal axis, concentration of bound XWnt8-AP. The fit of the data to a binary binding model gave K_D values of 16 and 18 nM for WIF-1 and WIF-1-IgG, respectively.

members of the Frizzled family (J.-C.H., A.R., P.M.S. and J.N., unpublished results). This XWnt8-Myc-conditioned medium, an analogous conditioned medium containing Wg, and a control conditioned medium from untransfected S2 cells were used in the following experiments.

WIF-1-IgG and an IgG fusion containing only the WIF domain (WD-IgG) co-precipitated Wg and XWnt8-Myc, with XWnt8-Myc being more efficient (Fig. 4c, d). The moderately efficient precipitation of Wg by hWIF-1-IgG (~10% co-precipitation; Fig. 4c) contrasts to the efficient inhibition by hWIF-1 and WIF-1-IgG of the Armadillo response to added Wg in clone-8 cells (Fig. 4b). This apparent discrepancy is explained if only a fraction of secreted Wg protein is active and can bind hWIF-1, WIF-1-IgG and the receptor on clone-8 cells. Binding seems to be specific and at least partially reversible, because a 20-fold excess of hWIF-1 can compete efficiently with WIF-1-IgG for binding to Wg or to XWnt8-Myc, and can displace 50% of Wg or 90% of XWnt8-Myc from a preformed complex with WIF-1-IgG in a 2-hour incubation at 4 °C (Fig. 4e).

Figure 4f shows that the binding efficiency of XWnt8-Myc for hWIF-1 is comparable to that of a fusion protein between the CRD of *Drosophila* Frizzled-2 and IgG (Dfz2CRD-IgG). In the presence of a 40-fold excess of soluble mouse sFRP-3 (msFRP-3), there was less co-precipitation of XWnt8-Myc with Dfz2CRD-IgG but the co-precipitation of XWnt8-Myc with WIF-1-IgG was unaffected. Consistent with this, XWnt8-Myc co-precipitated with WIF-1-IgG several-fold more efficiently than it did with an IgG fusion protein carrying the CRD of msFRP-3 (data not shown). By contrast, a 40-fold excess of soluble hWIF-1 significantly decreased the co-precipitation of XWnt8-Myc with Dfz2CRD-IgG and with WIF-1-IgG.

To quantify the binding affinity of hWIF-1 for XWnt8, a fusion of XWnt8-Myc and human placental alkaline phosphatase (XWnt8-AP) was produced in S2 cells and used for binding to hWIF-1, WIF-

1-IgG, or IgG coated onto 96-well plates. After 24 hours incubation at 4 °C, we detected little or no binding to IgG, and the hyperbolic binding curves for hWIF-1 and WIF-1-IgG were almost identical (Fig. 4g, left). A linear Scatchard plot of the binding data (Fig. 4g, right) is indicative of a simple non-cooperative interaction having a K_D of 16 nM.

These results with soluble secreted proteins indicate that at least two Wnt proteins, Wg and XWnt8-Myc, can form a specific, high-affinity, non-covalent complex *in vitro* with hWIF-1, and that hWIF-1 and Dfz2CRD bind to sites on XWnt8-Myc that either overlap or show negative cooperativity. These experiments do not address the possibility that additional proteins derived from the S2-conditioned medium might be present in our Wnt/hWIF-1 or Wnt/CRD complexes.

These results indicate that Wnt proteins can interact with structurally different extracellular inhibitors, as has been shown for the bone morphogenetic proteins^{8–10}. These extracellular inhibitors may fine-tune the spatial and temporal domains of Wnt effects or may act to transport Wnt proteins across the intercellular space, releasing them at the appropriate location¹¹. The presence of EGF-repeat domains in WIF-1 and a heparin-binding domain in sFRPs (ref. 12 and A.R., J.-C.H. and J.N., unpublished results) suggest that the spatial distribution of these Wnt antagonists is tightly controlled. A further level of complexity can be envisaged if, as our results indicate, various antagonists differ in their relative affinities for different Wnt proteins. The biological importance of such a system of Wnt antagonists could be the creation of spatial and temporal patterns of Wnt activity that are more complex than those allowed by transcriptional control of Wnt gene expression alone. □

Methods

Antibodies. Rabbit anti-WIF-domain antibodies were raised against a fusion protein composed of the T7 gene-10 protein fused to amino acids 29–168 of hWIF-1, and affinity-purified using an *E. coli* maltose-binding protein fused to the same WIF-1 segment. Rabbit anti-Wg antibodies were prepared by immunization with a fusion protein composed of amino acids 235–367 of Wg fused to glutathione-S-transferase (GST) and affinity-purified as described.

Protein production from transfected mammalian cells. The WIF-1-IgG fusion contains the entire hWIF-1 open reading frame, except for the C-terminal seven residues, upstream of the hinge region of the human IgG heavy-chain gene¹³. The WD-IgG fusion contains the first 180 residues of hWIF-1, a spacer of three glycines, and IgG. The Dfz2CRD-IgG fusion contains the first 271 residues of *Drosophila* Frizzled-2 upstream of IgG. hWIF-1 and the various IgG fusion proteins were produced in transiently transfected 293 cells. Serum-free conditioned medium was concentrated 20-fold by ultrafiltration. Control conditioned medium was prepared from untransfected 293 cells and similarly concentrated. Soluble msFRP-3 was purified from conditioned medium of transfected Chinese hamster ovary (CHO) cells that had been selected for methotrexate resistance following co-amplification of the msFRP-3 expression plasmid and a co-transfected dihydrofolate reductase plasmid. msFRP-3 in serum-free conditioned medium was purified to apparent homogeneity by heparin-Sepharose affinity chromatography.

Production of Wg, XWnt8-Myc and XWnt8-AP by S2 cells. Conditioned medium containing Wg was produced by heat-shocking S2 cells stably transfected with Wg cDNA under the control of a heat-shock promoter as described⁷. Conditioned medium containing XWnt8-Myc or XWnt8-AP was produced in stably transfected S2 cells by using a metallothionein promoter and incubating at 25 °C for 24 h in serum-free DES medium (Invitrogen) containing 0.5 mM CuSO₄. The Wg- or XWnt8-Myc-containing conditioned medium was centrifuged at 100,000g at 4 °C for 1 h to remove aggregates. Control media were produced by heat shock or CuSO₄ induction of untransfected S2 cells. Conditioned medium containing Wg and control conditioned medium for the Wg experiments were concentrated 20-fold by ultrafiltration. The XWnt8-Myc protein contains a C-terminal Myc-epitope tag (EQKLISEEDL) inserted between amino acids 339 and 340 (ref. 5); the XWnt8-AP construct contains alkaline phosphatase fused immediately beyond this Myc tag.

In situ hybridization. Whole-mount *in situ* hybridization was done as described¹⁴.

Table 1 WIF-1 inhibits axis induction by WNTs and promotes axis induction by Chordin in *Xenopus*

	<i>n</i>	Complete secondary axis (%)	Partial secondary axis (%)	Normal (%)
XWnt8 (2 pg)	41	88	10	2
XWnt8 (2 pg) + WIF-1 (20 pg)	28	0	81	19
XWnt8 (2 pg) + WIF-1 (200 pg)	37	0	68	32
XWnt8 (2 pg) + WIF-1 (1 ng)	59	0	27	73
XWnt8 (10 pg)	24	100	0	0
XWnt8 (10 pg) + WIF-1 (200 pg)	30	7	93	0
XWnt8 (10 pg) + WIF-1 (1 ng)	36	0	50	50
XWnt8 (2 pg) + WD (200 pg)	40	0	95	5
XWnt8 (2 pg) + WD (1 ng)	40	0	10	90
XWnt8 (2 pg) + β -gal (3 × 200 pg)*	17	50	50	0
XWnt8 (2 pg) + WIF-1 (3 × 100 pg)*	29	0	70	30
Wingless (2 pg)	38	92	8	0
Wingless (2 pg) + WIF-1 (20 pg)	22	91	9	0
Wingless (2 pg) + WIF-1 (200 pg)	21	86	14	0
Wingless (2 pg) + WIF-1 (1 ng)	41	0	70	30
XWnt3A (2 pg)	44	70	25	5
XWnt3A (2 pg) + WIF-1 (20 pg)	17	76	18	6
XWnt3A (2 pg) + WIF-1 (200 pg)	61	69	31	0
XWnt3A (2 pg) + WIF-1 (1 ng)	52	29	65	6
Siainos (20 pg)	36	100	0	0
Siainos (20 pg) + WIF-1 (200 pg/1 ng)	72	100	0	0
Chordin (200 pg)	35	0	95	5
Chordin (200 pg) + WIF-1 (200 pg)	41	100†	0	0
Chordin (200 pg) + Frzb (200 pg)	27	20‡	80	0
Chordin (200 pg) + WD (2–12 ng)	112	80‡	20	0

For Wnt inhibition experiments, embryos were injected in ventral blastomeres with the indicated RNAs, and secondary axes were scored at the tailbud stage. Duplicated axes were scored as complete when showing cement gland and eyes, and as partial when lacking both features.

* β -galactosidase or WIF-1 RNAs were injected at the 32-cell stage in three blastomeres adjacent to the blastomere that received XWnt8 RNA.

† 60% of these embryos had two separate eyes in the secondary head.

‡ All secondary heads were cyclopic.

Xenopus embryo and oocyte microinjection. hWIF-1 was expressed from the pCS2⁺ expression vector. RNA synthesis and microinjection into *Xenopus* embryos have been described¹⁵.

Armadillo stabilization assays. *Drosophila* clone-8 cells, seeded one day earlier and grown to 80% confluence, were incubated with control or Wg-containing conditioned medium from S2 cells. Before incubation with clone-8 cells, the S2 conditioned medium (0.4 ml) was pre-incubated for 25 min at 4°C with 0.4 ml DMEM/F-12 medium from transfected or control 293 cells. After 3 h at 25°C, clone-8 cells were collected, washed in 1 × PBS, 5 mM EDTA, and lysed in 80 µl hypotonic buffer (10 mM Tris, pH 7.5, 0.2 mM MgCl₂) containing protease inhibitors. After addition of 20 µl 1.25 M sucrose, a membrane-free cytoplasmic fraction was prepared by centrifugation at 100,000g for 30 min at 4°C, resolved by SDS-PAGE, immunoblotted and analysed for Armadillo (mAb N2-7A1; ref. 16), actin (Amersham) and HSP-70 (Sigma).

Solution binding assay for Wg and XWnt8-Myc. 200 µl 293-cell conditioned medium containing WIF-1-IgG, WD-IgG, or IgG (each adjusted by ultrafiltration to 60 nM) was incubated with protein A-Sepharose beads at 4°C for 1 h, after which the beads were washed 3 times with PBS and then incubated with 400 µl Wg or XWnt8-Myc conditioned medium at 4°C for 2 h. The beads were separated from unbound material by low-speed centrifugation and washed five times with PBS. Co-precipitates were analysed by SDS-PAGE and immunoblotted with affinity-purified rabbit anti-Wg antibodies or anti-Myc mAb 9E10 (ref. 17).

Quantitative binding of XWnt8-AP and hWIF-1. Conditioned medium (100 µl) containing WIF-1 (10 µg ml⁻¹), WIF-1-IgG (5 µg ml⁻¹), or IgG (4 µg ml⁻¹) was used to coat 96-well plate at 4°C overnight, followed by incubation at 4°C for 4 h with 200 µl 2 mg ml⁻¹ BSA in binding buffer (Hank's balanced salt, 20 mM HEPES, pH 7.0). 150 µl XWnt8-AP diluted in 2 mg ml⁻¹ BSA in binding buffer was applied to each well and incubated at 4°C for 24 h. After 5 washes with 200 µl each of binding buffer, bound XWnt8-AP was quantified by measuring alkaline phosphatase activity spectrophotometrically. A plot of alkaline phosphatase activity, representing the concentration of bound XWnt8-AP relative to the total concentration of XWnt8, was fitted to the simple binary binding model ($A + B \leftrightarrow A \cdot B$).

Received 17 December 1998; accepted 26 January 1999.

- Wodarz, A. & Nusse, R. Mechanisms of Wnt signaling in development. *Annu. Rev. Cell Dev. Biol.* **14**, 59–88 (1998).
- Moon, R. T., Brown, J. D. & Torres, M. WNTs modulate cell fate and behavior during vertebrate development. *Trends Genet.* **13**, 157–162 (1997).
- Glinka, A. et al. Dickkopf-1 is a member of a new family of secreted proteins and functions in head induction. *Nature* **391**, 357–362 (1998).
- Hoppler, S., Brown, J. D. & Moon, R. T. Expression of a dominant-negative Wnt blocks induction of MyoD in *Xenopus* embryos. *Genes Dev.* **10**, 2805–2817 (1996).
- Christian, J. L. & Moon, R. T. Interactions between XWnt-8 and Spemann organizer signaling pathways generate dorsoventral pattern in the embryonic mesoderm of *Xenopus*. *Genes Dev.* **7**, 13–28 (1993).
- Moon, R. T. et al. Dissecting Wnt signaling pathways and Wnt-sensitive developmental processes through transient misexpression analyses in embryos of *Xenopus laevis*. *Development* (suppl.), 85–94 (1993).
- van Leeuwen, F., Harryman Samos, C. & Nusse, R. Biological activity of soluble Wingless protein in cultured *Drosophila* imaginal disc cells. *Nature* **368**, 342–344 (1994).
- Piccolo, S., Sasai, Y., Lu, B. & DeRobertis, E. M. Dorsoventral patterning in *Xenopus*: inhibition of ventral signals by direct binding of chordin to BMP-4. *Cell* **86**, 589–598 (1996).
- Zimmerman, L. B., De Jesus-Escobar, J. M. & Harland, R. M. The Spemann organizer signal noggin binds and inactivates bone morphogenetic protein 4. *Cell* **86**, 599–606 (1996).
- Hsu, D. R., Economides, A. N., Wang, X., Eimon, P. M. & Harland, R. M. The *Xenopus* dorsaling factor gremlin identifies a novel family of secreted proteins that antagonize BMP activities. *Mol. Cell* **1**, 673–683 (1998).
- Cadigan, K. M., Fish, M. P., Rulifson, E. J. & Nusse, R. Wingless repression of *Drosophila* frizzled2 expression shapes the Wingless morphogen gradient. *Cell* **93**, 767–777 (1998).
- Finch, P. W. et al. Purification and molecular cloning of a secreted frizzled-related antagonist of Wnt signaling. *Proc. Natl Acad. Sci. USA* **94**, 6770–6775 (1997).
- Aruffo, A., Stamenkovic, I., Melnick, M., Underhill, C. B. & Seed, B. CD44 is the principal cell surface receptor for hyaluronate. *Cell* **61**, 1303–1313 (1990).
- Rebagliati, M. R., Toyoma, R., Haffner, P. & Dawid, I. B. Cyclops encodes a nodal-related factor involved in midline signaling. *Proc. Natl Acad. Sci. USA* **95**, 9932–9937 (1998).
- He, X., Saint-Jeanet, J.-P., Woodgett, J. R., Varmus, H. E. & Dawid, I. B. Glycogen synthase kinase-3 and dorsoventral patterning in *Xenopus* embryos. *Nature* **374**, 617–622 (1995).
- Peifer, M., Orsolic, S., Sweeton, D. & Wieschaus, E. A role for the *Drosophila* segment polarity gene *armadillo* in cell adhesion and cytoskeletal integrity during oogenesis. *Development* **118**, 1191–1207 (1993).
- Evan, G. I., Lewis, G. K., Ramsay, G. & Bishop, M. J. Isolation of monoclonal antibodies specific for human c-myc proto-oncogene product. *Mol. Cell. Biol.* **5**, 3610–3616 (1985).

Acknowledgements. We thank R. Moon for Xwnt8-Myc cDNA, J. Flanagan for the alkaline phosphatase plasmid, J. Corden for the mouse RNA polymerase II clone, B. Appel, L. Roman and D. Grunwald for cDNA libraries, and P. Bhanot and I. Munoz-Sanjuan for comments on the manuscript. Supported by the Howard Hughes Medical Institute (J.-C.H., A.R., P.M.S., C.H.S., R.N., J.N.).

Correspondence and requests for materials should be addressed to J.N. (e-mail: jnathans@jhmi.edu). Genbank accession numbers for WIF-1 are: human, AF122922; mouse, AF122923; *Xenopus*, AF122924; zebrafish, AF122925.

A capsaicin-receptor homologue with a high threshold for noxious heat

Michael J. Caterina*, Tobias A. Rosen*†, Makoto Tominaga*†, Anthony J. Brake* & David Julius*

* Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94143-0450, USA

† These authors contributed equally to this work

Pain-producing heat is detected by several classes of nociceptive sensory neuron that differ in their thermal response thresholds^{1–3}. The cloned capsaicin receptor, also known as the vanilloid receptor subtype 1 (VR1), is a heat-gated ion channel that has been proposed to mediate responses of small-diameter sensory neurons to moderate (43°C) thermal stimuli^{4,5}. VR1 is also activated by protons, indicating that it may participate in the detection of noxious thermal and chemical stimuli *in vivo*. Here we identify a structurally related receptor, VRL-1, that does not respond to capsaicin, acid or moderate heat. Instead, VRL-1 is activated by high temperatures, with a threshold of ~52°C. Within sensory ganglia, VRL-1 is most prominently expressed by a subset of medium- to large-diameter neurons, making it a candidate receptor for transducing high-threshold heat responses in this class of cells. VRL-1 transcripts are not restricted to the sensory nervous system, indicating that this channel may be activated by stimuli other than heat. We propose that responses to noxious heat involve these related, but distinct, ion-channel subtypes that together detect a range of stimulus intensities.

To identify new proteins involved in the detection of noxious stimuli by sensory neurons, we searched the GenBank database for sequences related to VR1. Most of the expressed sequence tag (EST) sequences identified appeared to encode human and mouse orthologues of the same protein, which we named vanilloid-receptor-like protein 1 (VRL-1). Using this information we isolated full-length VRL-1 complementary DNAs from rat brain and the CCRF-CEM human myeloid cell line (Fig. 1). These clones encode proteins of 761 amino acids and 764 amino acids, respectively, which share 78.4% identity and 86.2% similarity with one another. By comparison, rat and human VRL-1 are roughly 49% identical and 66% similar to rat VR1.

Like VR1, VRL-1 is predicted to contain six transmembrane domains, a putative pore-loop region, a cytoplasmic amino terminus with three ankyrin-repeat domains, and a cytoplasmic carboxy terminus. This overall architecture is characteristic of a family of ion channels defined by the transient receptor potential (TRP) and TRP-like (TRPL) channels of the *Drosophila* phototransduction pathway and which is now known to include vertebrate and nematode homologues^{6–8}. Of the TRP-related sequences reported so far, the one that resembles VR1 and VRL-1 most closely is that of the OSM-9 protein, which is required for osmosensation and the detection of some odorants in *Caenorhabditis elegans*⁸. Even OSM-9, however, is only 23% identical to rat VR1 and 24% identical to rat VRL-1, indicating that these latter molecules may form a distinct subgroup within this growing family of proteins.

VR1 is activated by capsaicin (the main pungent ingredient in 'hot' chilli peppers) and by its very potent analogue, resiniferatoxin (from *Euphorbia* plants). These vanilloid compounds elicit pain by evoking non-selective cationic currents in small-diameter nociceptive neurons (nociceptors)^{4,9–11}. Even at very high concentrations, however, neither capsaicin (Fig. 2a, d) nor resiniferatoxin (not shown) evoked currents in *Xenopus* oocytes or HEK293 human embryonic kidney cells expressing VRL-1. Thus, VRL-1 does not

rVRL-1	MTSASS-PPAFRLETSDG-DEEGNAE-----VNGKQEP---PPMESPFQREDRNS	46
hVRL-1	MTSPSS-SPVFRLETLDG-GQEDGSE-----ADRGLDFGSGLPPEMQSQFGEDRKF	50
rVR1	MEQRASLDSESESPQENSCLDPPDRPNCKPPVKPHIFTTRSRTRLFGKGDSEE---ASPLDCPYEEGLAS	72
rVRL-1	SPQIKVNLNFIKRPPKN--TSAPSQQE-----PDR-FDRDRLFSSVSRGVPEELTGLLEYLRWNSKYLTDSA	110
hVRL-1	APQIRVNLNRYRK-----TGASQPD-----PNR-FDRDRLFNAVSRGVPEELAGLPEYLSKTSKYLTDSA	109
rVR1	CPIITVSSVLTIQRPDGPASVRPSSQDSVSAGEKPPRLYDRRSIFDAVAQSNCELESLLPFLQRSKKRLTDSA	147
----- A1 -----		
rVRL-1	YTEGSTGKTCLMKAVLNLDQGVNACIMPLLLQIDKDSGNPKPLVNAQCTDEFYQGHSAHIAIEKRSQCCKVLLVE	185
hVRL-1	YTEGSTGKTCLMKAVLNLDQGVNACIMPLLLQIDKDSGNPKPLVNAQCTDDYRGHSALHIAIEKRSQCCKVLLVE	184
rVR1	FKDPETGKTCLLKMALNLHNGQNDITALLLDVARKTDSLKQFVNASYTDSYKGTALHIAIERNNMTLVTLTLLVE	222
----- A2 -----		
rVRL-1	NGADVHLRACGRFFQKHQG-TCFYFGELPLSLAACTQKQWVVTYLLNPHQPASLEATDSLGNVTLHALVMIADN	259
hVRL-1	NGANVHARACGRFFQKHQG-TCFYFGELPLSLAACTQKQWVVTYLLNPHQPASLEATDSLGNVTLHALVMIADN	258
rVR1	NGADVQAANGDFFKKTKGRPGFYFGELPLSLAACTNQLAIVKFLQNSWQPADISARDSVGNVTLHALVEVADN	297
----- A3 -----		
rVRL-1	SPENSALVIHMYDGLLQMGARLCPTVQLEEISNHQGLTPLKLAKEGKIEIFRHILQREFSGP-YQPLSRKFTFW	333
hVRL-1	SAENIALVTSMYDGLLQMGARLCPTVQLEEISNHQGLTPLKLAKEGKIEIFRHILQREFSGP-YQPLSRKFTFW	331
rVR1	TVDNTKFTSMYNEILILGAKLHPTLKLEETNRKGLTPLALAASSGKIGVLAYILQREIHEPECRHLSRKFTFW	372
----- TM1 -----		
rVRL-1	CYGPVVRVSLYDLSSVDSWEKNSVLEIIAFHCK-SPNRHRMVLEPLNKLQEKWDRLVSR-FFNFACVLYVMFI	406
hVRL-1	CYGPVVRVSLYDLASVDSCEENSVLEIIAFHCK-SPNRHRMVLEPLNKLQAKWDLIPK-FFNLNCLNLYMFI	404
rVR1	AYGPVHSLYDLSDIDTCEKNSVLEIAYSSSETPNRHRMMLVLEPLNRLQDKWDRFVKRIYFNFFVYCYMFI	447
----- TM2 -----		
rVRL-1	FTVVAYHQPSLDQPAIPSSKATFGESMLLLGHILILLGGIYLLGLWYFWRRLFIWISFMSDYFEILFLQAL	481
hVRL-1	FTAVAYHQPTLKKQAAPHLKAEVGNMMLTGHILILLGGIYLLGLWYFWRRLFIWISFIDSYFEILFLQAL	479
rVR1	FTAAAYRYPVEGLPPY-KLKNVTGDFRVGTGILSVSGGVYFFFRGIQYFLQRRPSLKSFLVDSYSEILFFVQSL	521
----- TM4 -----		
rVRL-1	LSQVLRFMETEWYPLLLVSLVLGWLNLNLYYTRGFQHTGIYSVMIQKVLRLDRLFLVYLVFLFGFAVALV	556
hVRL-1	VSQVLCFLAIEWYPLLLVSLVLGWLNLNLYYTRGFQHTGIYSVMIQKVLRLDRLFLVYLVFLFGFAVALV	554
rVR1	FMLVSVVLYFSQRKEYVASMVF AM T M QM A E M C MF ST V	596
----- Pore Loop -----		
rVRL-1	SLSREARSPKAPEDNNSVTVEQPTVGQEEEP--APYRS DASLELFKFTIGMGELAFQELRFRGVVLLLLLAY	629
hVRL-1	SLSQEAWRPEAPTGPNATESVQPMGQDEGNGAQRGILEASLELFKFTIGMGELAFQELHFRGMVLLLLLAY	629
rVR1	TLIEDGKNNSLPMS---TPHKCRGSACKP-GNSYNSLYSTCLELFKFTIGMGDLFEFTENYDFKAVFIILLAY	666
----- TM6 -----		
rVRL-1	VLLTYVLLLNMLIALMSETVNVHADNWSIWLKQKASVLEMENGYWWCRKKHREGRLLVGTRGDGTPDERWC	704
hVRL-1	VLLTYVLLLNMLIALMSETVNSVATDSWSIWLKQKASVLEMENGYWWCR-KKQKAGVMTVGTGPDGSPDERWC	703
rVR1	VILTYVLLLNMLIALMGETVNKIAQESKN R TI DT KSFLKCMRKAFRSGKLLQVGFTPDGKDDYRWC	741
rVRL-1	FRVEEVNWAWEKTLPTLSEDP-SGPGITGNKNKPTSK---PGK-----NSASEEDHLP----LQVLQ	759
hVRL-1	FRVEEVNWAWEQTLPTLSEDP-SGAGVPTLENVPLAS--PPKEDE-----DGASEENYVP----VQLLQ	762
rVR1	FRVDEVNWTWNTNVGINEDPGNCEGVKRTLSFSLRSGRVSGRNWKNFALVPLLRDASTDRHATQEEVQLKH	816
rVRL-1	SP	761
hVRL-1	SN	764
rVR1	YTGSLKPEDAIEVFKDSMVPGEK	838

Figure 1 Alignment of cDNA sequences of rat VRL-1, human VRL-1 and rat VR1. Ankyrin-repeat domains (dashed black lines), putative transmembrane (TM)

domains (solid black lines) and the pore-loop region (dotted black line) are indicated. Identical residues are shaded in grey.

appear to be a vanilloid receptor. VR1 can also be activated by low pH (<6.0) or by thermal stimuli sufficient to cause pain in mammals (>43 °C)^{4,5}. We therefore investigated whether protons or heat could evoke currents in HEK293 cells or oocytes expressing VRL-1. Bath acidification to pH 4.0 did not produce responses in either system, and neither did temperature increases in the low-to-moderate noxious range (<50 °C). We did, however, observe pronounced currents when bath temperatures were raised from 22 °C to >53 °C (Fig. 2a, d). These responses, seen in 15 out of 20 batches of VRL1-injected oocytes and in ~70% of VRL1-transfected HEK293 cells, were markedly larger than the background heat-evoked currents observed in control water-injected oocytes or vector-transfected HEK293 cells, indicating that they may be mediated by VRL-1. No such currents were observed in oocytes injected with complementary RNA encoding the 5HT₃ serotonin receptor, or in HEK293 cells expressing the P2X₂ ATP receptor (not shown).

Having identified a stimulus for VRL-1, we asked whether this

channel shares pharmacological or electrophysiological properties in common with VR1. Capsazepine (10 µM), a competitive antagonist that blocks vanilloid-, proton-, or heat-evoked responses in VR1-expressing cells^{4,5,12}, had no effect on heat-evoked currents in VRL1-expressing oocytes (not shown). In contrast, the non-competitive VR1 antagonist ruthenium red^{4,5,13} inhibited heat-evoked VRL-1 responses significantly (90.1 ± 5% at 10 µM, mean ± s.d.) and dose dependently (half-maximal inhibitory concentration = 0.62 µM, Fig. 2b, c). The current-voltage relationship of heat-evoked VRL-1 responses in transfected HEK293 cells showed dual rectification, with a more prominent outward component (Fig. 2e). Although VR1-mediated currents exhibit only outward rectification at both whole-cell and single-channel levels^{4,5}, rectification of VRL-1 more closely resembles that seen for the *Drosophila* TRP channel¹⁴. Currents mediated by VRL-1 are cationic, with a permeability sequence Ca²⁺ > Mg²⁺ > Na⁺ ≈ Cs⁺ ≈ K⁺, and relative permeability ratios (Fig. 2f) similar to those for heat-evoked currents in VR1-expressing cells³. This functional similarity may reflect the

sequence conservation within the putative pore regions of these two proteins (Fig. 1).

Single-unit recordings from peripheral nerves have shown that heat-sensitive nociceptors can be subdivided into medium- and high-threshold classes, depending on whether they fire when temperatures exceed $\sim 43^\circ\text{C}$ or $\sim 53^\circ\text{C}$, respectively^{1–3}. The threshold for VR1 activation is $\sim 43^\circ\text{C}$ in transfected HEK293 cells (Fig. 3a, middle tracing)⁵. In contrast, heat-evoked currents in

VRL1-expressing HEK293 cells did not develop until the bath temperature exceeded 50°C (Fig. 3a, top tracing). A more detailed thermal response profile generated at a series of plateaux bath temperatures yielded a threshold value for VRL-1 activation of $50\text{--}52^\circ\text{C}$, compared with the lower activation threshold of $\sim 44^\circ\text{C}$ for VR1 (Fig. 3b). When VR1 and VRL-1 were co-expressed in HEK293 cells, a bimodal heat-evoked current was observed that appeared to represent the addition of a desensitizing medium-threshold VR1 response and a high-threshold VRL-1 response (Fig. 3a, bottom tracing). These findings indicate that VR1 and VRL-1 channels can account for responses of medium- and high-threshold nociceptors, respectively, to heat.

When temperature ramps were applied to oocytes expressing VR1 or VRL-1, initial heat-evoked currents developed at about 45°C or 53°C , respectively (Fig. 3c, d), again showing that these two channel subtypes have distinct thresholds for heat activation. These continuous thermal-response profiles also illustrate the extremely steep temperature dependence of the currents once threshold is reached. In the oocyte system, the thermal sensitivities of both VR1 and VRL-1 increased with repeated heat applications (Figs 2b, 3d); by the second or third stimulus, either channel could be activated by temperatures as low as 40°C . This pattern of sensitization resembles

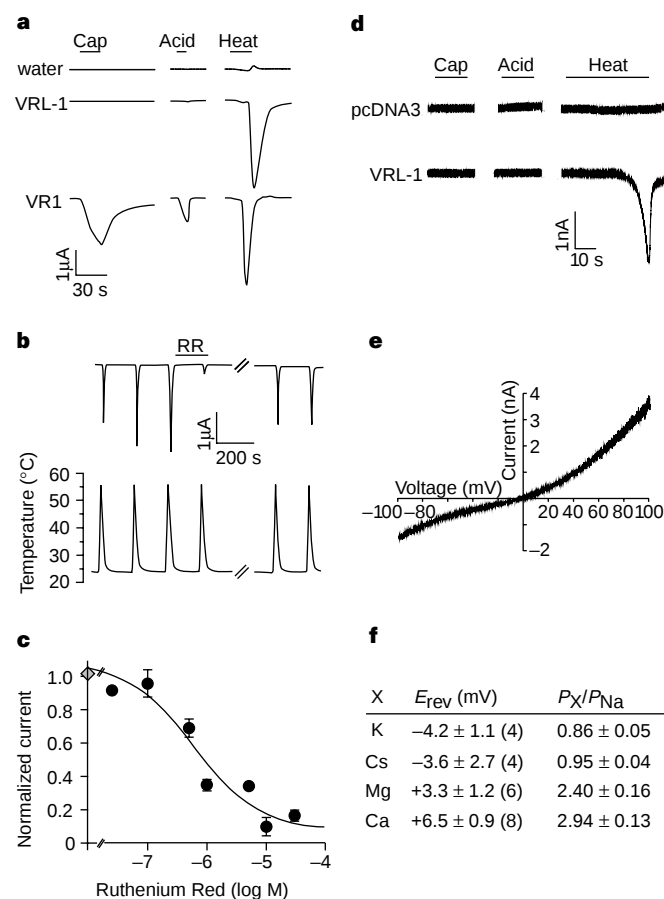


Figure 2 VRL-1 is activated by heat, but not by vanilloids or protons. **a**, Voltage-clamp analysis of oocytes injected with water (top), VRL-1 cRNA (middle), or VR1 cRNA (bottom). Oocytes were exposed to capsaicin ($1\text{ }\mu\text{M}$), protons (pH 4) or heat (23°C to 55°C in 15 s). Duration of heat stimuli varied by ~ 3 s between experiments, but peak temperatures were within $\pm 0.4^\circ\text{C}$ of each other. Capsaicin ($100\text{ }\mu\text{M}$), resiniferatoxin ($10\text{ }\mu\text{M}$), ATP ($100\text{ }\mu\text{M}$), acetylcholine ($300\text{ }\mu\text{M}$), histamine ($10\text{ }\mu\text{M}$), substance P ($1\text{ }\mu\text{M}$), bradykinin ($1\text{ }\mu\text{M}$) and serotonin ($10\text{ }\mu\text{M}$) did not evoke currents in VRL1-injected oocytes (not shown). **b**, Ruthenium red (RR, $10\text{ }\mu\text{M}$) reversibly inhibits heat-evoked currents in oocytes expressing VRL-1. The wash-out period (starting parallel lines) was 10 min. **c**, Dose dependence of RR-mediated inhibition of heat-evoked (54°C) currents in oocytes expressing VRL-1. Current magnitudes were normalized to that of the third heat response in a given oocyte. Mean values ($\pm$ s.e.m., $n = 4\text{--}6$ oocytes per point) were fitted to the Hill equation. In oocytes not treated with RR, the fourth heat-evoked response was the same as the third (1.01 ± 0.03 -fold, diamond). **d**, Whole-cell voltage-clamp analysis of HEK293 cells transfected with vector (pcDNA3) or VRL-1 cDNA. Cells were successively exposed to capsaicin ($100\text{ }\mu\text{M}$), protons (pH 4) and heat (22°C to 54°C in 45 s). Resiniferatoxin ($1\text{ }\mu\text{M}$) evoked no current in VRL1-expressing cells (not shown). **e**, Current-voltage relationship for heat-evoked currents in VRL1-transfected HEK293 cells. A voltage ramp was applied from -100 mV to $+100$ mV in 500 ms. The reversal potential (E_{rev}), obtained with caesium aspartate pipette solution, was -3.1 ± 0.5 mV ($n = 10$). **f**, Relative cationic permeabilities ($P_{\text{X}}/P_{\text{Na}}$) of heat-evoked currents in VRL1-transfected HEK293 cells. Values shown are means $\pm$ s.e.m. (number of independent cells).

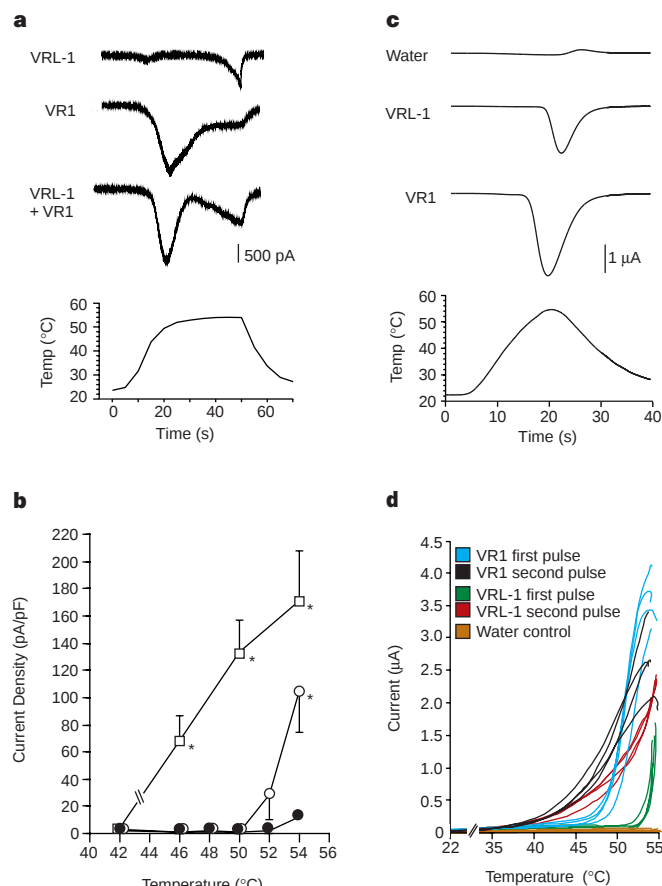


Figure 3 VRL-1 exhibits a higher activation threshold than VR1. **a**, Currents evoked by temperature ramps in HEK293 cells expressing VRL-1 (top), VR1 (middle), or both (bottom). Bath temperature is plotted below (final temperature 54°C). **b**, Temperature-response profile of heat-evoked currents in HEK293 cells transfected with pcDNA3 (filled circles), VR1 (squares) or VRL-1 (open circles) plasmids. Asterisks indicate significant differences from pcDNA3 ($P < 0.01$, unpaired Student's t -test) (pcDNA3, $n = 5$; VR1, $n = 7$; VRL-1, $n = 24$). **c**, Currents evoked by a temperature ramp in oocytes injected with water or VRL-1 or VR1 cRNA. **d**, Response profiles of four oocytes exposed to two successive heat ramps, as in **c**.

that exhibited by some nociceptors following an initial exposure to noxious heat². We also subjected VRL1-expressing oocytes to two consecutive subthreshold heating cycles (23 °C to 50 °C), followed by a cycle spanning 23 °C to 56 °C. In this case, currents were observed only in the third cycle, after the temperature exceeded 53 °C (not shown), showing that the higher threshold observed for VRL-1 does not reflect simply a latent response to a lower threshold temperature. Moreover, these results show that subthreshold temperatures do not sensitize VRL-1.

Thermally responsive nociceptors differ not only in their activation thresholds, but also with respect to size and myelination pattern¹⁵. C-polymodal nociceptors are small-diameter, unmyelinated neurons that respond to heat with a threshold of ~43 °C. Aδ mechano- and heat-sensitive (AMH) neurons are medium- to large-diameter, lightly myelinated neurons that fall into two groups: type I AMHs have a response threshold of ~53 °C, and type II AMHs are activated at ~43 °C (refs 1, 3, 15). We therefore used VR1- and VRL1-specific antisera to determine whether the distribution of these two receptors within sensory ganglia shows a correlation with these cell types. In rat dorsal root ganglia (DRG), VR1 expression was observed in 40% of all neurons (157 out of 397), predominantly in a population with small cell bodies ($19.2 \pm 0.3 \mu\text{m}$, mean diameter $\pm$ s.e.m.; Figs 4a, 5 top). This observation is consistent with previous histological and functional studies^{4,5,16–18}.

Intense VRL-1 immunoreactivity was observed in 16.4% of all DRG neurons (152/925) (arrowheads in Fig. 4b), and an analysis of size distribution showed that these cells were of medium to large diameter ($29.0 \pm 0.6 \mu\text{m}$, mean diameter $\pm$ s.e.m.; Fig. 5, bottom). More moderate VRL-1 immunoreactivity was seen in neurons with very large diameters, where staining was confined to the plasma membrane (arrows in Fig. 4b). Similar expression profiles were observed in trigeminal ganglia (not shown). Very few of the intensely VRL1-positive cells stained with lectin IB4 (1/58, Fig. 4c) or substance P antibody (3/57, not shown), markers for two populations of small-diameter nociceptive neurons¹⁹. However, about one-third of the intensely VRL1-positive cells (41/114) co-

stained with antiserum against CGRP (arrows in Fig. 4e, f), consistent with the expression pattern reported for rat Aδ neurons²⁰.

Distinct distributions of VR1 and VRL-1 were also observed in primary cultures of rat DRG neurons (not shown). Roughly 80% of VRL1-immunoreactive cells (85/105) in these cultures stained with the anti-neurofilament antibody RT97 (Fig. 4d), a marker for myelinated neurons^{21,22}. Within the spinal cord, intense VRL-1 immunoreactivity was observed in Lissauer's tract and in the dorsal horn (Fig. 4g). Dorsal horn staining was most intense within the marginal zone and lamina I (Fig. 4g, arrowhead), as well as in a band spanning the boarder between inner lamina II and lamina III (Fig. 4g, arrow). This pattern, confirmed by double labelling with IB4 (not shown), resembles that formed by central terminals of Aδ primary afferents²³, although a contribution from intrinsic dorsal horn neurons cannot be excluded. VRL-1 immunoreactivity was also seen in the dorsal columns (consistent with the lower-level immunoreactivity observed in large-diameter DRG neurons), in motoneurons and in astrocytes (not shown).

These histological results, together with the temperature-response data, indicate that robust expression of VRL-1 by a group of Aδ neurons may underlie the characteristically high thermal threshold ascribed to type I AMH nociceptors. Expression of VR1 by C-nociceptors could account for the moderate thermal threshold exhibited by these neurons. This proposal is in agreement with the results of a recent electrophysiological study²⁴ that showed that cultured sensory neurons fall into two classes with regard to heat sensitivity: low-threshold (45 °C), capsaicin-responsive cells with small diameters (~19 μm) and high-threshold (51 °C), capsaicin-insensitive cells with larger diameters (~26 μm). Finally, we found that VRL-1 messenger RNA is expressed in tissues other than sensory ganglia and spinal cord, including lung, spleen, intestine, and brain (most subregions) (Fig. 4h, and data not shown). In these locations, VRL-1 may be activated by stimuli other than noxious heat.

Genetic analyses in invertebrate systems have placed TRP and related proteins downstream of G-protein-coupled signalling cascades, most notably in pathways involving the hydrolysis of mem-

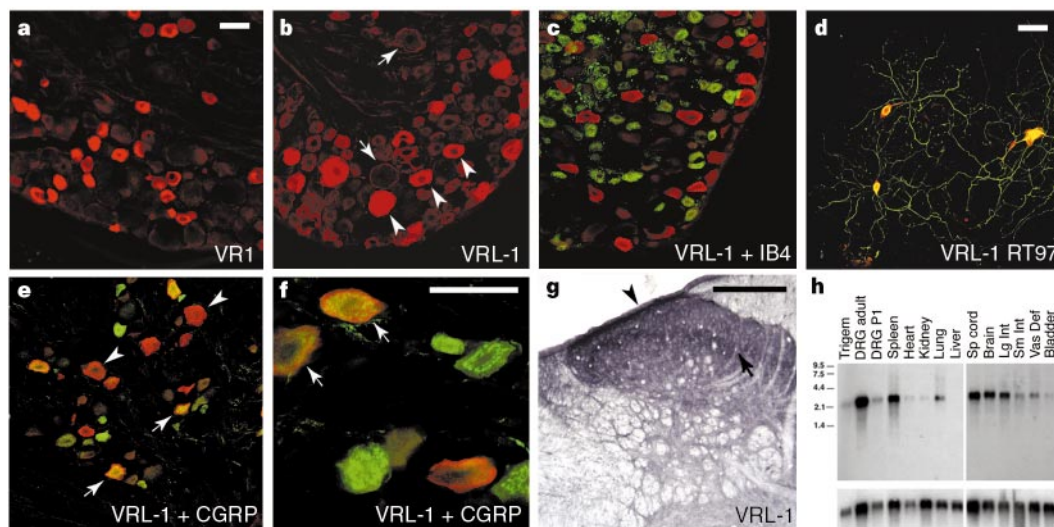


Figure 4 VRL-1 is highly expressed in a subset of medium- to large-diameter sensory neurons. Rat DRG sections (15 μm) were stained with: **a**, anti-VR1; **b**, anti-VRL1; **c**, anti-VRL1 (red) plus FITC-lectin IB4 (green); **e**, **f**, anti-VRL1 (red) and anti-CGRP (green) antibodies. **d**, Cultured sensory neurons stained with anti-VRL1 (red) and RT97 anti-neurofilament (green) antibodies. Double staining appears yellow. **g**, Spinal cord sections stained with anti-VRL1 antibody. **h**, Northern blot analysis of VRL-1 mRNA expression in rat tissues. Blots were probed with VRL-1 cDNA (top), and with rat cyclophilin cDNA (bottom) to normalize for loading. DRG

P1 indicates neonatal DRG. Molecular-size markers (in kilobases) are shown at the left. Scale bar in **a** (50 μm) also applies to **b**, **c**, **e**. Scale bars in **d**, **f**, **g** represent 100, 50, and 200 μm, respectively. In **b**, arrowheads indicate intense staining of medium- to large-diameter neurons, and arrows indicate staining of neurons with very large diameters. In **e**, **f**, arrows and arrowheads indicate double- (yellow) and single-positive cells, respectively. VRL-1 peptide, but not control peptide, greatly diminished VRL-1 staining (not shown).

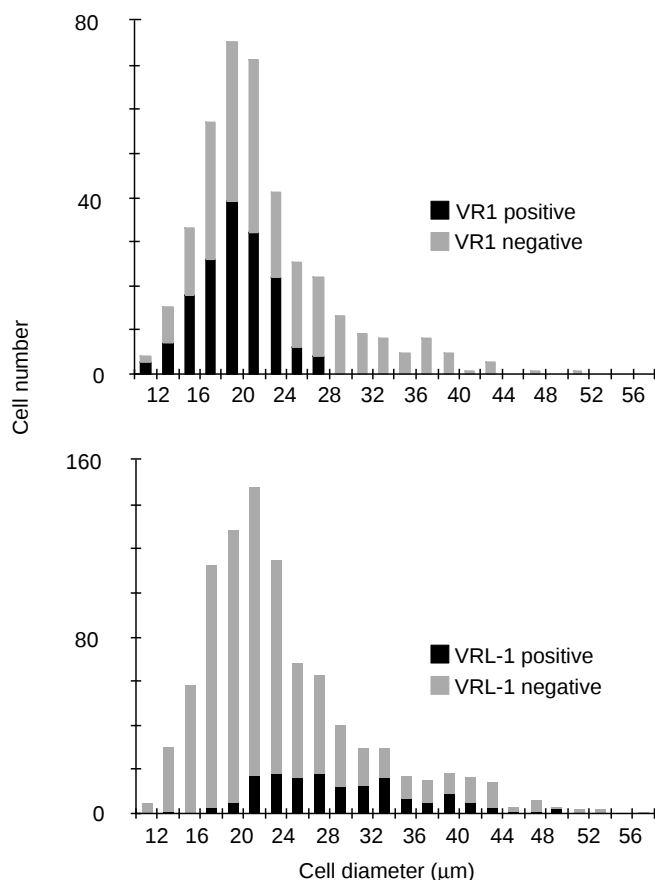


Figure 5 VR1 and VRL-1 proteins show distinct expression patterns among DRG neurons of different size classes. Only intensely labelled cells were counted as positive.

brane lipids^{25,26}. We found VRL-1 transcripts in many non-neuronal tissues. As these tissues are unlikely to encounter temperatures above 50 °C, other physiological stimuli might regulate the channel at these sites, either directly or through the release of (perhaps G-protein-mediated) second messengers. The thermal-response characteristics of VRL-1 might also be controlled by such mechanisms. Nevertheless, the observation that VR1 and VRL-1 exhibit distinct thresholds for activation by heat indicates that their absolute thermal sensitivities may be governed, at least in part, by their primary sequences.

Individual members of receptor-protein families are often tuned to detect a stimulus over a discrete window of intensities. Together, an ensemble of such receptor subtypes can recognize a physiological stimulus over a wide dynamic range. Perhaps the best example of such a strategy is provided by the vertebrate visual system²⁷, where different opsin subtypes are tuned to discrete frequency ranges of the electromagnetic spectrum. By analogy, our findings indicate that related ion channels may account for thermal responsiveness over a range of noxious temperatures. In addition to its ability to sense noxious heat, the nervous system is responsive to a broad range of noxious and innocuous temperatures, from warm to cold. It will therefore be interesting to determine the extent to which vanilloid receptor-like subtypes contribute to this sensory repertoire. □

Methods

Molecular biology. Polymerase chain reaction (PCR) primers (5'-GACCA GCAAGTACCTCAC-3' and 5'-CTCCCATGCAGCCAGTTTACTTCCACCC TGAAGCACCAGCGCTCA-3') were designed from VR1-related EST sequences and used to amplify a 1,985-nucleotide fragment from neonatal rat brain as

described²⁸. A full-length rat VRL-1 cDNA containing a 2,283-nucleotide open reading frame with 5' and 3' untranslated regions of 329 and 105 nucleotides, respectively, was isolated from a rat brain cDNA plasmid library prepared as described²⁸. Two other primers (5'-ATAAGAATGCGGCCGCGAGAGGTCCT GGCTGGAC-3' and 5'-GGAAGATATCAGAGCCAGCAGTATGTGTTG-3') were used to amplify a full-length human VRL-1 cDNA from CCRF-CEM cells (a gift from J. Imboden).

Mammalian-cell electrophysiology. HEK293 cells were transfected using calcium phosphate with VR1 cDNA (diluted 1:200 with pcDNA3) or undiluted VRL-1 cDNA. In co-transfection experiments, VR1, VRL-1 and pcDNA3 plasmids were mixed at a ratio of 1:10:29 (total 12 μg DNA per 35-mm dish). For control experiments using P2X₂, cells were transfected with 1 μg of plasmid DNA using lipofectamine (Gibco). Whole-cell patch-clamp recordings were carried out as described⁴. Standard bath solution for whole-cell recordings contained (in mM): 140 NaCl, 5 KCl, 2 MgCl₂, 2 CaCl₂, 10 HEPES, 10 glucose, pH 7.4 (adjusted with NaOH). In some experiments, bath solution was buffered to pH 4.0 with 10 mM MES. For monovalent-cation substitution experiments, bath solution was changed to (in mM): 140 NaCl (or KCl or CsCl), 10 glucose and 10 HEPES (adjusted to pH 7.4 with NaOH, KOH or CsOH, respectively) and reversal potential was measured using voltage ramps (-100 to +40 mV in 500 ms). For divalent-cation substitution experiments, bath solution was changed to (in mM): 110 MgCl₂ (or CaCl₂), 2 Mg(OH)₂ (or Ca(OH)₂), 10 glucose, 10 HEPES, pH 7.4 (adjusted with HCl). Standard pipette solution contained (in mM): 140 CsCl (or 130 caesium aspartate and 10 NaCl for current-voltage analysis), 5 EGTA, 10 HEPES, pH 7.4 (adjusted with CsOH). Pipette solution for cation-substitution experiments contained (in mM): 140 NaCl, 10 HEPES, 5 EGTA, pH 7.4 (adjusted with NaOH). Bath temperature was raised by perfusion with preheated solutions and monitored using a thermocouple (Omega). For each heating protocol, averaged temperature profiles were established during bath perfusion ($n \geq 5$) before cellular recording. Liquid-junction potentials (8.6–9.6 mV) were measured in separate experiments and corrections made for membrane potentials. Relative permeability ratios (P_X/P_{Na}) were calculated as described⁴. VRL1-transfected cells were defined as heat responsive if heat-evoked currents were greater than two standard deviations above the mean current observed in vector-transfected cells.

Oocyte electrophysiology. *Xenopus* oocytes were injected with 5–10 ng VR1, 25 ng VRL-1, or 25 ng 5HT₃ cRNA, as described⁴. Two-electrode voltage-clamp analysis ($E_h = -40$ mV) was carried out 7–16 days post-injection. Recording solution contained (in mM): 90 NaCl, 1.0 KCl, 2.4 NaHCO₃, 0.1 BaCl₂, 1.0 MgCl₂, 10 HEPES, pH 7.6, unless otherwise indicated. Bath temperature was increased using an in-line solution heater (SH27A, Warner Instruments, Inc) and a timed relay. Chamber temperature was monitored (accuracy ± 0.2 °C) with a Super MCJ thermocouple (Omega) placed within 2 mm of the oocyte. VRL1-injected oocyte batches were defined as heat responsive if more than half of the cells exhibited currents $\geq$ twofold those observed in water-injected oocytes. Roughly half of all VRL1-injected oocyte batches exhibited currents $\geq$ fivefold those exhibited by controls.

Localization of VRL-1 expression. Northern blot analysis was done using a ³²P-labelled fragment (nucleotides 381–2,299) of rat VRL-1 cDNA as described²⁸. Rabbits were immunized (HTI, Inc.) with keyhole limpet haemocyanin-conjugated peptide corresponding to the C terminus of rat VRL-1 plus an N-terminal cysteine (peptide sequence CKNSASEEDHPLQLVQSP). Affinity purification of the antiserum was done after coupling the peptide to amine-reactive resin (Pierce). The antiserum specifically recognized a band of relative molecular mass $\sim 80,000$ ($M_r \sim 80$ K) in whole-cell extracts prepared from HEK293 cells transfected with rat VRL-1 cDNA (not shown). For peptide competition, diluted antibody solutions (5 μg ml⁻¹) were preincubated (4 °C, 60 min) with 5–20 μg ml⁻¹ VRL1-antigenic peptide or control peptide (CEDAEVFKDSMVPGEK).

Fixed frozen sections from adult rat sensory ganglia (15 μm) and spinal cord (35 μm) were prepared as described⁵. Immunofluorescence was performed as described⁵ using rabbit anti-VRL1 (0.83 μg ml⁻¹), rabbit anti-VR1 (0.5 μg ml⁻¹), fluorescein isothiocyanate (FITC)-conjugated lectin IB4 (8 μg ml⁻¹, Sigma) guinea pig anti-substance P (1:12,000; Peninsula), guinea pig anti-CGRP (1:200, Peninsula), mouse monoclonal RT97 antibodies (10 μg ml⁻¹, Boehringer), and/or mouse anti-NeuN (1:3,000; Chemicon),

followed by Cy3-labelled goat anti-rabbit (1:700; Jackson Immunoresearch), Cy2-labelled donkey anti-guinea-pig (1:100; Jackson) and/or FITC-labelled horse anti-mouse (1:500; Vector) antibodies. For size-distribution studies, immunofluorescently stained sections double-labelled with anti-NeuN²⁹ were analysed using NIH image software. Glucose oxidase/nickel-enhanced diaminobenzidine immunostaining of spinal cord sections was performed using anti-VRL1 (0.83 $\mu\text{g ml}^{-1}$) as described⁵.

Primary cultures prepared from adult rat DRG³⁰ were incubated overnight (37 °C, 5% CO₂) in medium containing nerve growth factor (100 ng ml⁻¹) and fixed with 10% formalin in 0.1 M phosphate buffer. For size-distribution studies, cells were stained with anti-VRL1 or anti-VRL1 IgG (166 ng ml⁻¹), detected with diaminobenzidine-peroxidase (Vector), and analysed as above.

Received 9 November 1998; accepted 24 February 1999.

- Dubner, R., Price, D. D., Beitel, R. E. & Hu, J. W. in *Pain in the Trigeminal Region* (eds Anderson, D. J. & Matthews, B.) 57–66 (Elsevier, Amsterdam, 1977).
- Campbell, J. N. & Meyer, R. A. in *Spinal Afferent Processing* (ed. Yaksh, T. L.) 59–81 (Plenum, New York, 1986).
- Leem, J. W., Willis, W. D. & Chung, J. M. Cutaneous sensory receptors in the rat foot. *J. Neurophysiol.* **69**, 1684–1699 (1993).
- Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
- Tominaga, M. *et al.* The cloned capsaicin receptor integrates multiple pain-producing stimuli. *Neuron* **21**, 531–543 (1998).
- Montell, C. & Rubin, G. M. Molecular characterization of the *Drosophila trp* locus: a putative integral membrane protein required for phototransduction. *Neuron* **2**, 1313–1323 (1989).
- Zhu, X. *et al.* *trp*, a novel mammalian gene family essential for agonist-activated capacitative Ca²⁺ entry. *Cell* **85**, 661–671 (1996).
- Colbert, H. A., Smith, T. L. & Bargmann, C. I. Osm9, a novel protein with structural similarity to ion channels, is required for olfaction, mechanosensation and olfactory adaptation in *Caenorhabditis elegans*. *J. Neurosci.* **17**, 8259–8269 (1997).
- Marsh, S. J., Stansfield, C. E., Brown, D. A., Davey, R. & McCarthy, D. The mechanism of action of capsaicin on sensory C-type neurons and their axons *in vitro*. *Neuroscience* **23**, 275–289 (1987).
- Bevan, S. & Szolcsanyi, J. Sensory neuron-specific actions of capsaicin: mechanisms and applications. *Trends Pharmacol. Sci.* **11**, 330–333 (1990).
- Winter, J., Walpole, C. S., Bevan, S. & James, I. F. Characterization of resiniferatoxin binding and capsaicin sensitivity in adult rat dorsal root ganglia. *Neuroscience* **57**, 747–757 (1993).
- Bevan, S. *et al.* Capsazepine: a competitive antagonist of the sensory neuron excitant capsaicin. *Br. J. Pharmacol.* **107**, 544–552 (1992).
- Dray, A., Forbes, C. A. & Burgess, G. M. Ruthenium red blocks the capsaicin-induced increase in intracellular calcium and activation of membrane currents in sensory neurones as well as the activation of peripheral nociceptors *in vitro*. *Neurosci. Lett.* **110**, 52–59 (1990).
- Reuss, H., Mojet, M. H., Chyb, S. & Hardie, R. *In vivo* analysis of the *Drosophila* light-sensitive channels, TRP and TRPL. *Neuron* **19**, 1249–1259 (1997).
- Meyer, R. A., Campbell, J. N. & Raja, S. in *Textbook of Pain* (eds Wall, P. D. & Melzack, R.) 13–44 (Churchill Livingstone, Edinburgh, 1994).
- Helliwell, R. J. A. *et al.* Capsaicin sensitivity is associated with the expression of vanilloid (capsaicin) receptor (VR1) mRNA in adult rat sensory ganglia. *Neurosci. Lett.* **250**, 177–180 (1998).
- Jancso, G., Kiraly, E. & Jancso-Gabor, A. Pharmacologically induced selective degeneration of chemosensitive primary sensory neurons. *Nature* **270**, 741–743 (1977).
- Wood, J. N. *et al.* Capsaicin-induced ion fluxes in dorsal root ganglion cells in culture. *J. Neurosci.* **8**, 3208–3220 (1988).
- Snider, W. D. & McMahon, S. B. Tackling pain at the source: new ideas about nociceptors. *Neuron* **20**, 629–632 (1998).
- McCarthy, P. W. & Lawson, S. N. Differing action potential shapes in rat dorsal root ganglion neurones related to their substance P and calcitonin gene-related peptide immunoreactivity. *J. Comp. Neurol.* **388**, 541–549 (1997).
- Lawson, S. N. & Waddell, P. J. The antibody RT97 distinguishes between sensory cell bodies with myelinated and unmyelinated peripheral processes in the rat. *J. Physiol.* **371**, 59P (1985).
- Wood, J. N. & Anderson, B. H. Monoclonal antibodies to mammalian neurofilaments. *Biosci. Rep.* **1**, 263–268 (1981).
- Light, A. & Perl, E. R. Re-examination of the dorsal root projection to the spinal dorsal horn including observations on the differential termination of coarse and fine fibers. *J. Comp. Neurol.* **186**, 117–132 (1979).
- Nagy, I. & Rang, H. Noxious heat activates all capsaicin-sensitive and also a subpopulation of capsaicin-insensitive dorsal root ganglion neurons. *Neuroscience* **88**, 995–997 (1999).
- Hardie, R. C. & Mink, B. Novel Ca²⁺ channels underlying transduction in *Drosophila* photoreceptors: implications for phosphoinositide-mediated Ca²⁺ mobilization. *Trends Neurosci.* **16**, 371–376 (1993).
- Scott, K. & Zucker, C. TRP, TRPL and trouble in photoreceptor cells. *Curr. Opin. Neurobiol.* **8**, 383–388 (1998).
- Nathans, J. Molecular biology of visual pigments. *Annu. Rev. Neurosci.* **10**, 163–194 (1987).
- Brake, A., Wagenbach, M. J. & Julius, D. New structural motif for ligand-gated ion channels defined by an ionotropic ATP receptor. *Nature* **371**, 519–523 (1994).
- Todd, A. J., Spike, R. C. & Polgar, E. A quantitative study of neurons which express neurokinin-1 or somatostatin sst2a receptor in rat spinal cord dorsal horn. *Neuroscience* **85**, 459–473 (1998).
- Reichling, D. B. & Levine, J. D. Heat transduction in rat sensory neurons by calcium-dependent activation of a cation channel. *Proc. Natl Acad. Sci. USA* **94**, 7006–7011 (1997).

Acknowledgements. We thank M. Bland for assistance with immunolocalization; L. England, J. Onuffer and members of the Basbaum laboratory for advice regarding primary neuronal culture, immunolocalization, and affinity purification methods; and A. Basbaum, H. Chuang, L. England, H. Ingraham and S. Jordt for comments on the manuscript. M.J.C. is an American Cancer Society postdoctoral fellow and NARSAD young investigator; M.T. is a Comroe Fellow of the UCSF Cardiovascular Research Institute. This work was supported by grants from NIGMS and NIDR.

Correspondence and requests for materials should be addressed to D.J. (e-mail: julius@socrates.ucsf.edu). The Genbank accession numbers for rat and human VRL-1 cDNAs are AF129113 and AF129112, respectively.

Prisoner's dilemma in an RNA virus

Paul E. Turner* & Lin Chao*

Department of Biology, University of Maryland, College Park, Maryland 20742, USA

The evolution of competitive interactions among viruses¹ was studied in the RNA phage $\phi 6$ at high and low multiplicities of infection (that is, at high and low ratios of infecting phage to host cells). At high multiplicities, many phage infect and reproduce in the same host cell, whereas at low multiplicities the viruses reproduce mainly as clones. An unexpected result of this study¹ was that phage grown at high rates of co-infection increased in fitness initially, but then evolved lowered fitness. Here we show that the fitness of the high-multiplicity phage relative to their ancestors generates a pay-off matrix conforming to the prisoner's dilemma strategy of game theory^{2,3}. In this strategy, defection (selfishness) evolves, despite the greater fitness pay-off that would result if all players were to cooperate. Viral cooperation and defection can be defined as, respectively, the manufacturing and sequestering of diffusible (shared) intracellular products. Because the low-multiplicity phage did not evolve lowered fitness, we attribute the evolution of selfishness to the lack of clonal structure and the mixing of unrelated genotypes at high multiplicity^{4–6}.

Evolutionary game theory predicts the outcome of pairwise contests between players that use conflicting strategies³. If a population of individuals playing one strategy cannot be invaded by mutants playing any other strategy, the former becomes the evolutionarily stable strategy. But if no single strategy is able to resist invaders, a polymorphic equilibrium with both strategies ensues. Whether a single strategy or a polymorphic equilibrium evolves depends on the fitness pay-offs to the players. In a contest between the strategies of cooperation and defection, the resulting 2×2 pay-off matrix is described by three variables (Fig. 1a). When a pair of cooperators interact, their individual fitness has a value of one. When a cooperator and a defector are paired, the cooperator is exploited and its fitness is decreased by s_1 , whereas the defector benefits and its fitness is increased by s_2 . When two defectors interact, they suffer from not having anyone to exploit and pay a cost c . If $(1 - c) < (1 - s_1)$ a polymorphic equilibrium results. If $(1 - c) > (1 - s_1)$ the population evolves to be 100% defectors. The latter is evolutionarily paradoxical, and termed the prisoner's dilemma, because a population composed of defectors has a lower fitness than one containing only cooperators. Although the prisoner's dilemma is clearly of evolutionary importance, it is difficult in most biological systems to measure the pay-off values associated with it^{7–9} and most examples are theoretical¹⁰.

Viral evolution offers a unique opportunity to study the prisoner's dilemma because co-infection of the same host cell by more than one virus creates conflicts^{1,11–13} similar to those assumed in game theory, and ancestral genotypes can often be retrieved for reconstructing the pay-off matrix. The manufacture of diffusible and hence shared intracellular products by viruses co-infecting the same cell allows for the strategies of cooperation and defection. A viral genotype that synthesizes larger quantities of products is effectively a cooperator. In contrast, a genotype that synthesizes less but specializes in sequestering a larger share of the products is a defector. The often observed evolution of defective interfering particles¹⁴ in viruses supports this interpretation. The particles are coded by viral RNAs that have lost most or all of their protein-coding sequences,

* Present addresses: Instituto Cavanilles de Biodiversidad y Biología Evolutiva and Departamento de Genética, Universidad de Valencia, 46100 Valencia, Spain (P.E.T.); and Department of Biology, University of California San Diego, La Jolla, California 92093-0116, USA (L.C.).

a		b	
Cooperate	Cooperate	$\phi 6$	$\phi H2$
	Defect	$\phi 6$	$\phi H2$
Defect	Cooperate	1	0.65
	Defect	$1+s_2$	1.99
			0.83

Figure 1 Expected and observed fitness values for a game in which opponents use conflicting strategies of cooperation and defection. **a**, Generalized pay-off matrix where entries represent the fitness to an individual adopting the strategy on the left, if the opponent adopts the strategy above. Defectors gain a fitness advantage ($1+s_2$) that allows them to invade a population of cooperators. If the cost of defection is too strong, $(1-c) < (1-s_1)$, cooperators may also invade and the two strategies are driven to a stable polymorphism. The prisoner's dilemma occurs if it always pays to be selfish, $(1-c) > (1-s_1)$; defection sweeps through the population despite the greater fitness pay-off that would result had all individuals cooperated. **b**, Realized pay-off matrix for the evolved high MOI phage $\phi H2$ relative to its ancestor $\phi 6$ reveals evolution of an evolutionarily stable strategy conforming to the prisoner's dilemma. Observed fitness values indicate that the lowest pay-off is to the ancestor during co-infection. Thus, the evolutionarily stable strategy is to defect even though a higher pay-off occurs when phage cooperate.

and are thus unable to reproduce in the absence of complete viruses. This inability defines the defective nature of defective interfering particles but they are also functionally defectors (in the game theory sense) because they have evolved a variety of mechanisms to gain an intracellular fitness advantage in the presence of complete viruses. Their smaller size allows some particles to replicate faster, whereas others skip unnecessary transcription steps or are preferentially processed because they feature extra sequences recognized by encapsidation and replication enzymes¹⁴⁻¹⁶. By providing the necessary enzymes for defective interfering particles, complete viruses act functionally as cooperators.

Because a population composed only of defective interfering particles is unable to reproduce and $(1-c)$ equals zero, the evolution of the particles results in a polymorphic equilibrium and does not constitute a case of the prisoner's dilemma. Thus it becomes important to determine whether complete (non-defective interfering) viruses can also evolve quantitatively different strategies on a continuum of cooperation and defection, and whether any of the resulting strategies conform to the prisoner's dilemma. A previous study allowed replicate populations of $\phi 6$ to evolve at high and low multiplicities-of-infection (MOI)¹. Data showed that phage cultured at high MOI (but not low MOI) gain an added advantage during co-infection, indicating that these viruses evolved a defection strategy for intracellular competition. Further results indicated that in environments where high MOI phage become fixed, and hence intracellular competition with other genotypes is removed, these viruses exhibit evolution of lowered fitness. Because the evolution of lowered fitness in a population of defectors is expected from the prisoner's dilemma, we have investigated further whether game theory could be used to interpret this surprising result.

Although the prisoner's dilemma leads to fixation (a population playing a single strategy), its effects should still be frequency dependent because defectors gain their greatest fitness advantage when rare and interacting primarily with cooperators (Fig. 1a). Thus, we first sought evidence of whether fitness of the evolved high MOI phage was dependent on their initial frequency in competition. We isolated two clones ($\phi H1$ and $\phi H2$) at random from separate populations evolved at high MOI. The fitness of each clone was measured relative to the ancestor, $\phi 6$, at an MOI of five that reflects the high degree of co-infection experienced by $\phi H1$ and $\phi H2$ during their evolution. A host-range mutation¹⁷ was intro-

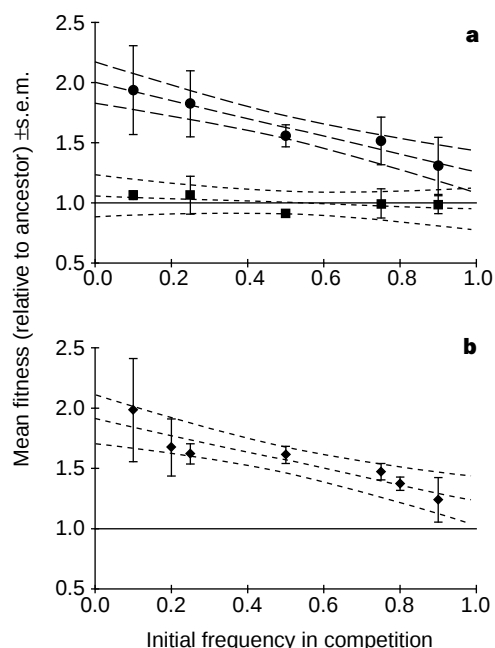


Figure 2 The fitness of derived high MOI phage relative to the ancestor is a decreasing function of initial frequency in competition. **a**, Evolved phage $\phi H2$ and a genetically marked clone of the ancestor, $\phi 6$, were competed at five initial frequencies of $\phi H2$ (0.1, 0.25, 0.5, 0.75, 0.9) at MOI = 5 with fourfold replication. Linear regression analysis shows that the fitness of $\phi H2$ (circles) is dependent upon its initial frequency in competition (slope = -0.7381 , $t_s = -8.117$, d.f. = 3, $P = 0.0039$), whereas a control that measures fitness of $\phi 6$ (squares) relative to the marked clone is independent of initial frequency (see Methods for statistics). Values are the means $\pm$ s.e.m. Dashed lines indicating the 95% confidence interval about each regression line are included to show that mean fitness of $\phi H2$ exceeds 1.0 at all initial frequencies. Phage $\phi H2$ is $\phi S2$ from a previous study¹. **b**, Experiments where a different evolved clone, $\phi H1$, and the marked ancestor were competed at seven initial frequencies of $\phi H1$ (0.1, 0.2, 0.25, 0.5, 0.75, 0.8, 0.9) at MOI = 5 with fourfold replication. Once again, linear regression analysis shows that the fitness of $\phi H1$ (diamonds) is dependent upon its initial frequency in competition (slope = -0.6789 , $t_s = -5.128$, d.f. = 5, $P = 0.0037$), and exceeds 1.0 at all initial frequencies.

duced into $\phi 6$ to serve as a genetic marker that differentiates the ancestor during fitness assays. The results (Fig. 2) clearly show that the fitness of $\phi H1$ and $\phi H2$ is a decreasing function of each clone's initial frequency, whereas controls involving $\phi 6$ and the marked ancestor demonstrate that frequency-dependence is not driven by the host-range marker. More importantly, these results also demonstrate that for frequencies approaching 1.0 the mean fitness of $\phi H1$ and $\phi H2$ exceeds unity. A fitness value equal to 1.0 would have implied that the fitness of the high MOI phage was equal to that of $\phi 6$. Rather, the fitness of $\phi H1$ and $\phi H2$ exceeds 1.0 at all initial frequencies despite the negative effect of frequency on fitness, demonstrating that $\phi H1$ and $\phi H2$ will achieve fixation as required by the prisoner's dilemma.

The support for the prisoner's dilemma in Fig. 2 is best illustrated when these data are used to estimate s_1 , s_2 and c in our pay-off matrix (Fig. 1a). Because the fitness data for $\phi H1$ and $\phi H2$ are virtually identical, we chose to explore the prisoner's dilemma hypothesis further using only $\phi H2$ (Fig. 2a). Fitness was measured at a high MOI, where most phage reproduce in co-infected cells. At low initial frequencies of $\phi H2$, pure co-infections containing only $\phi H2$ are rare and the fitness of $\phi H2$ is primarily determined by mixed co-infections that also contain $\phi 6$. Thus, the fitness of $\phi H2$ at low frequencies is equal to $(1+s_2)$. It follows that $\phi 6$ is very abundant at low frequencies of $\phi H2$ and most co-infection events will occur through the ancestor alone. If the fitness of $\phi 6$ in a pure

co-infection is defined to be one, then the y -intercept of the regression line in Fig. 2a equals $(1 + s_2)/1$, which provides the estimate of $(1 + s_2) = 1.99$. By the same logic, the fitnesses of ϕ H2 and ϕ 6 at high initial frequencies of ϕ H2 are $(1 - c)$ and $(1 - s_1)$, whereby the ratio $(1 - c)/(1 - s_1)$ is estimated by the extrapolated fitness value in Fig. 2a at a frequency of 1.0. Because the extrapolated value of 1.28 is greater than 1.0, $(1 - c) > (1 - s_1)$ as required by the prisoner's dilemma.

However, to demonstrate simply that $(1 - c) > (1 - s_1)$ does not provide definitive support for the prisoner's dilemma. Rather, to complete the pay-off matrix, either $(1 - c)$ or $(1 - s_1)$ must be estimated separately and we devised an experiment that measures $(1 - c)$. Adsorption is the step in phage infection that involves attachment and entry of viruses into the bacterial cell¹⁸. We modified our fitness assay so that each phage was allowed to adsorb separately at a high MOI, and then mixed the adsorbed phage (each at a frequency of 0.5) just before the fitness assay. This approach differs from our previous experiments (Fig. 2) in which the phage were mixed before adsorption. Thus, whereas previous assays contained cells co-infected by both ϕ H2 and ϕ 6, the present assay contains only cells infected with either ϕ H2 or ϕ 6 genotypes. Without co-infection, the fitness of ϕ H2 is $(1 - c)$ and that of ϕ 6 is 1 (Fig. 1a). Results showed $(1 - c)$ to be 0.830 ± 0.051 (mean $\pm$ s.e.m., $n = 10$), in which case $(1 - s_1) = 0.83/1.28 = 0.65$. Note that in our previous assays involving mixed infections the fitness of ϕ H2 is much higher at a frequency of 0.5 (compare with Fig. 2a).

Substitution of parameter estimates into our pay-off matrix allows us to reconstruct the evolution of viral fitness at high MOI (Fig. 1b). The defection strategy exhibited by ϕ H2 affords a large fitness advantage relative to the ancestor during co-infection, allowing this mutant to invade when rare. As ϕ H2 increases in frequency its intracellular interaction with other genotypes becomes more rare, causing its fitness relative to the ancestor (or other co-operators) to dip below 1.0. This idea is supported by previous data where intrahost competition between viruses was removed¹, and by our current estimates of $(1 - c)$. However, our data also show $(1 - c) > (1 - s_1)$, indicating the lowest fitness pay-off is to the ancestor when ϕ H2 is very common. Thus, it always pays to be selfish and the defection strategy shown by ϕ H2 is able to sweep through the population. Because all genotypes wind up in the lower right of the matrix despite its inherently low fitness pay-off, this clearly conforms to the prisoner's dilemma.

One way for evolving populations to escape the prisoner's dilemma is through kin selection (but see ref. 19 for an alternative mechanism). During intrahost reproduction, the relatedness of parasite progeny selects for different strategies to use limited host resources^{4,6,20,21}. In single infections, either competition is absent or selection favours the evolution of cooperation among closely related individuals. In contrast, mixing of unrelated genotypes favours defection to exploit resources selfishly. Therefore, evolution of parasite strategies during intrahost competition can be viewed as a problem of kin selection^{5,22}. We allowed the evolution of a single ancestral virus in two environments where degree of progeny relatedness differed¹. At high MOI, strong intracellular competition occurred between distantly related individuals, whereas at low MOI competition was either absent or occurred among closely related kin. Thus, our findings that evolution of phage at high MOI results in prisoner's dilemma can be attributed to the absence of clonal structure at high MOI. Lack of clonal structure leads to the evolution of selfishness, but the traditional view is that selfishness evolves through more rapid exploitation of host resources, generally resulting in the evolution of increased parasite virulence and often a concomitant shortened lifespan of the host^{23,24}. Our study shows an alternative pathway for selfishness to evolve, that is, viruses gain an advantage by sequestering gene products of co-infecting genotypes, which does not necessarily lead to an increase in virulence. Our data show that fitness trade-offs consistent with the prisoner's dilemma

and other game theoretic models readily evolve in biological systems as simple as viruses. Furthermore, the prisoner's dilemma provides a clear case for game theory to challenge the idea that natural selection should always lead to a fitness increase²⁵. □

Methods

Culture conditions and fitness assays. Bacterial cultures and phage lysates were grown, incubated and diluted at 25 °C using LC broth and agar²⁶. To measure fitness, a genotype was mixed at a defined ratio with a genetically marked clone. The mixture was then allowed 40 min adsorption to *Pseudomonas phaseolicola* at MOI = 5, where greater than 99% of phage are adsorbed by 40 min²⁷. Before cells had burst²⁷, the diluted mixture was plated on LC agar with a *P. phaseolicola* lawn, and then incubated for 24 h. The resulting ~500 plaques on the plate were then collected²⁶ and filtered (0.22 μ m Durapore, Millipore) to remove bacteria. Some fitness assays allowed separate adsorption for competitors. The marker used was a point mutation on the medium segment that allowed growth on the alternative host *P. pseudocaligenes*¹⁷. Ratio of phages in the starting mixture (R_0) and that in the collected lysate (R_1) were monitored by plating on mixed lawns of *P. phaseolicola* and *P. pseudocaligenes* (200:1 ratio), on which wild-type and host-range clones make turbid and clear plaques, respectively. The number of plaques per plate was kept at 500 by diluting the collected phage with LC broth. A maximum of 500 was chosen to minimize overlap, and thus genetic exchange, between plaques. Fitness (W) was defined as the relative change of the ratio wild-type:host-range, or $W = R_1/R_0$. Competitions between the ancestor and a marked clone (Fig. 2a) showed a small deleterious effect to the marker (mean W of ancestor = 1.0454 ± 0.041 s.e.m.; $n = 20$), that did not differ according to initial frequency of the ancestor (linear regression with slope = -0.1144 , $t_s = -1.202$, d.f. = 3, $P = 0.3156$). In addition, similar competitions at a frequency of 0.5 showed a marker effect (mean W of ancestor = 1.0598 ± 0.024 s.e.m.; $n = 12$) that did not differ according to the presence of co-adsorption (independent samples t -test with $t_s = 0.219$, d.f. = 10, $P = 0.831$). All fitness values reported were therefore scaled to adjust for the effects of the marker by setting the mean fitness of the ancestral clone to 1.0.

Received 23 December 1998; accepted 15 February 1999.

1. Turner, P. E. & Chao, L. Sex and the evolution of intrahost competition in RNA virus ϕ 6. *Genetics* **150**, 523–532 (1998).
2. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
3. Maynard-Smith, J. *Evolution and the Theory of Games* (Cambridge Univ. Press, Cambridge, 1982).
4. Hamilton, W. D. Altruism and related phenomena, mainly in social insects. *Annu. Rev. Ecol. Syst.* **3**, 193–232 (1972).
5. Frank, S. A. A kin selection model for the evolution of virulence. *Proc. R. Soc. Lond. B* **250**, 195–197 (1992).
6. Nowak, M. A. & May, R. M. Superinfection and the evolution of parasite virulence. *Proc. R. Soc. Lond. B* **255**, 81–89 (1994).
7. Craig, J. L. Are communal puke caught in a Prisoner's Dilemma? *Behav. Ecol. Sociobiol.* **14**, 147–150 (1984).
8. Fischer, E. A. Simultaneous hermaphroditism, Tit for Tat, and the evolutionary stability of social systems. *Ethol. Sociobiol.* **9**, 119–136 (1988).
9. Chao, L. Evolution of polyandry in a communal breeding system. *Behav. Ecol.* **8**, 668–674 (1997).
10. Dugatkin, L. A. *Cooperation Among Animals: an Evolutionary Perspective* (Oxford Univ. Press, Oxford, 1997).
11. von Magnus, P. Incomplete forms of influenza virus. *Adv. Virus Res.* **2**, 59–79 (1954).
12. Lewontin, R. C. The units of selection. *Annu. Rev. Ecol. Syst.* **1**, 1–18 (1970).
13. Bonhoeffer, S. & Nowak, M. A. Intra-host and inter-host selection: viral evolution of immune function impairment. *Proc. Natl Acad. Sci. USA* **91**, 8062–8066 (1994).
14. Huang, A. S. & Baltimore, D. Defective viral particles and viral disease processes. *Nature* **226**, 325–327 (1970).
15. Holland, J. *Fundamental Virology* 2nd edn (eds Fields, B. & Knipe, D.) 151–165 (Raven, New York, 1991).
16. Chao, L. *The Evolutionary Biology of Viruses* (ed. Morse, S. S.) 233–250 (Raven, New York, 1994).
17. Mindich, L., Cohen, J. & Weisburd, M. Isolation of nonsense suppressor mutants in *Pseudomonas*. *J. Bact.* **126**, 177–182 (1976).
18. Stent, G. *Molecular Biology of Bacterial Viruses* (Freeman, San Francisco, 1963).
19. Nowak, M. & May, R. Evolutionary games and spatial chaos. *Nature* **359**, 826–829 (1992).
20. Herre, E. A. Population structure and the evolution of virulence in nematode parasites of fig wasps. *Science* **259**, 1442–1445 (1993).
21. Frank, S. A. Models of parasite virulence. *Q. Rev. Biol.* **71**, 37–78 (1996).
22. Hamilton, W. D. The genetical evolution of social behavior: I and II. *J. Theor. Biol.* **7**, 1–52 (1964).
23. Bremermann, H. J. & Pickering, J. A game-theoretical model of parasite virulence. *J. Theor. Biol.* **100**, 411–426 (1983).
24. Knolle, H. Host density and the evolution of parasite virulence. *J. Theor. Biol.* **136**, 199–207 (1989).
25. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
26. Sinclair, J. F., Cohen, J. & Mindich, L. The isolation of suppressible nonsense mutants of bacteriophage ϕ 6. *Virology* **75**, 198–208 (1976).
27. Vidaver, K. A., Koski, R. K. & Van Etten, J. L. Bacteriophage ϕ 6: a lipid-containing virus of *Pseudomonas phaseolicola*. *J. Virol.* **11**, 799–805 (1973).

Acknowledgements. We thank C. Burch, K. Hanley, S. Lance, U. Mueller, J. Smale and G. Wilkinson for useful comments and enlightening discussion. This work was supported by fellowships from the NSF and University of Maryland to P.E.T.

Correspondence and requests for materials should be addressed to P.E.T. (e-mail: paul.e.turner@uv.es).